

Testing hypotheses via a mixture estimation model*

KANIAV KAMARY

Université Paris-Dauphine, CEREMADE

KERRIE MENGENSEN

Queensland University of Technology, Brisbane

CHRISTIAN P. ROBERT

Université Paris-Dauphine, CEREMADE, Dept. of Statistics, University of Warwick, and CREST, Paris

JUDITH ROUSSEAU

Université Paris-Dauphine, CEREMADE, and CREST, Paris

Abstract. We consider a novel paradigm for Bayesian testing of hypotheses and Bayesian model comparison. Our alternative to the traditional construction of Bayes factors or posterior probabilities of a model given the data is to consider the hypotheses or models under comparison as components of a mixture model. We therefore replace the original testing problem with an estimation one that focuses on the probability or weight of a given hypothesis within the mixture model. We analyse the sensitivity of the posterior distribution of the weights induced by different priors on these quantities. We stress that a major appeal in using this novel perspective is that generic improper priors are acceptable. Among other features, this allows for a resolution of the Lindley–Jeffreys paradox. For example, a reference Beta $\mathcal{B}(a_0, a_0)$ prior on the mixture weights can be used for the common problem of testing two contrasting hypotheses or models. In this case the sensitivity of the posterior estimates of the weights to the choice of a_0 vanishes as the sample size increases. We therefore advocate a default choice of $a_0 = 0.5$, derived from [Rousseau and Mengersen \(2011\)](#). Similar results apply for testing more than two alternatives. Another feature of this easily implemented alternative to the classical Bayesian solution is that the speeds of convergence of the posterior mean of the weight and of the corresponding posterior probability are quite similar.

Key words and phrases: testing statistical hypotheses, Bayesian analysis, mixture model, mixture estimation, improper prior, Beta prior,

*Kaniav Kamary, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France, kamary@ceremade.dauphine.fr, Kerrie Mengersen, QUT, Brisbane, QLD, Australia, k.mengersen@qut.edu.au, Christian P. Robert and Judith Rousseau, CEREMADE, Université Paris-Dauphine, 75775 Paris cedex 16, France, xian,rousseau@ceremade.dauphine.fr. Research partly supported by the Agence Nationale de la Recherche (ANR, 212, rue de Bercy 75012 Paris) through the 2012–2015 grant ANR-11-BS01-0010 “Calibration”, a 2010–2015 senior chair grant of Institut Universitaire de France and an Australian Research Council grant. Thanks to Phil O’Neill and Theo Kypriaos from the University of Nottingham for an interesting discussion about our different approaches to testing via mixture. We are also grateful to the participants to BAYSM’14 in Vienna and ISBA’14 in Cancun for their comments, and to Jean-Louis Foulley for pointing out the econometrics references.

Dirichlet prior, posterior probability, Bayes factor.

1. INTRODUCTION

Hypothesis testing is one of the central problems of statistical inference and a dramatically differentiating feature of classical and Bayesian paradigms (Neyman and Pearson, 1933; Berger and Sellke, 1987; Casella and Berger, 1987; Gigerenzer, 1991; Berger, 2003; Mayo and Cox, 2006; Gelman, 2008). However, the way in which hypothesis testing is conducted in a Bayesian framework is wide open to controversy and divergent opinions (Jeffreys, 1939; Bernardo, 1980; Berger, 1985; Aitkin, 1991; Berger and Jefferys, 1992; De Santis and Spezzaferri, 1997; Bayarri and Garcia-Donato, 2007; Christensen et al., 2011; Johnson and Rossell, 2010; Gelman et al., 2013a; Robert, 2014). In particular, the handling of non-informative Bayesian testing is mostly unresolved and has produced much debate; see, for example, the specific case of the Lindley or Jeffreys–Lindley paradox (Lindley, 1957; Shafer, 1982; DeGroot, 1982; Robert, 1993; Lad, 2003; Spanos, 2013; Sprenger, 2013; Robert, 2014).

Bayesian hypothesis testing can be considered as a model selection problem, which allows the comparison of several potential statistical models in order to identify the model that is most strongly supported by the data. There is a range of ways in which this problem can be interpreted. For instance, Christensen et al. (2011) consider it to be a decision issue that pertains to testing, while Robert (2001) express it as a model index estimation setting and Gelman et al. (2013a) disagree about the decision aspect. Notwithstanding this debate, the most common approaches to Bayesian hypothesis testing in practice are posterior probabilities of the model given the data (Robert, 2001), the Bayes factor (Jeffreys, 1939) and its approximations such as the Bayesian information criterion (BIC) and the Deviance information criterion (DIC) (Schwarz, 1978; Csiszár and Shields, 2000; Spiegelhalter et al., 2002; Plummer, 2008) and posterior predictive tools and their variants (Gelman et al., 2013a; Vehtari and Lampinen, 2002; Vehtari and Ojanen, 2012).

A standard Bayesian approach to testing (Berger, 1985; Robert, 2001) is to consider two families of models, one for each of the hypotheses under comparison,

$$\mathfrak{M}_1 : x \sim f_1(x|\theta_1), \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x \sim f_2(x|\theta_2), \theta_2 \in \Theta_2,$$

to associate with each of those models a prior distribution,

$$\theta_1 \sim \pi_1(\theta_1) \quad \text{and} \quad \theta_2 \sim \pi_2(\theta_2),$$

and to compute the marginal likelihoods

$$m_1(x) = \int_{\Theta_1} f_1(x|\theta_1) \pi_1(\theta_1) d\theta_1 \quad \text{and} \quad m_2(x) = \int_{\Theta_2} f_2(x|\theta_2) \pi_2(\theta_2) d\theta_2.$$

The hypotheses or models can then be compared either through the *Bayes factor* or the posterior probability, respectively:

$$\mathfrak{B}_{12} = \frac{m_1(x)}{m_2(x)}, \quad \mathbb{P}(\mathfrak{M}_1|x) = \frac{\omega_1 m_1(x)}{\omega_1 m_1(x) + \omega_2 m_2(x)}.$$

Note that the latter quantity depends on the prior weights ω_i of both models. Both testing and model selection are thus expressed as a comparison of models. The Bayesian decision step proceeds by comparing the Bayes factor \mathfrak{B}_{12} to the threshold value of one, or comparing the posterior probability $\mathbb{P}(\mathfrak{M}_1|x)$ to a bound derived from a 0–1 loss function (or a “golden” bound

like $\alpha = 0.05$ inspired from frequentist practice (Berger and Sellke, 1987; Berger et al., 1997, 1999; Berger, 2003; Ziliak and McCloskey, 2008). As a general rule, when comparing more than two models, the model with the highest posterior probability is selected. However, this rule is highly dependent on the prior modelling, even with large datasets, which makes it hard to promote as the default solution in practical studies.

Some well-documented difficulties with this traditional handling of Bayesian tests and Bayesian model choices via posterior probabilities are as follows.

- ✓ There is a tension between using these two approaches. Posterior probabilities are justified by a binary loss function but depend on prior weights. In contrast, Bayes factors eliminate this dependence but do not fully encompass the posterior distribution, unless the prior weights are integrated into the loss function (Berger, 1985; Robert, 2001).
- ✓ The interpretation of the strength of the Bayes factor in supporting a given hypothesis or model is delicate and somewhat arbitrary (Jeffreys, 1939; Dickey and Gunel, 1978; Kass and Raftery, 1995; Lavine and Schervish, 1999). This is mainly due to the fact that it is not a Bayesian decision rule unless, as above, the loss function is artificially modified to incorporate the prior weights.
- ✓ An analogous difficulty with posterior probabilities is a tendency to interpret them as p -values when in fact they only report through a marginal likelihood ratio the respective strengths of the data under each of the models (and of course nothing about the “truth” of either model).
- ✓ There is a long-lasting impact of the choice of the prior distributions on the parameter spaces of both models under comparison. This occurs despite the existence of an overall consistency proof for the Bayes factor (Berger et al., 2003; Rousseau, 2007; McVinish et al., 2009).
- ✓ Improper priors are prohibited, since they are not theoretically justified in most testing situations (DeGroot, 1970, 1973; Robert, 2001, 2014). This drawback, discussed in more detail below, has led to the proliferation of alternative *ad hoc* solutions, where the data are either used twice (Aitkin, 1991, 2010; Gelman et al., 2013b) or split in artificial ways (O’Hagan, 1995; Berger and Pericchi, 1996; Berger et al., 1998; Berger and Pericchi, 2001).
- ✓ The use of a binary (*accept* vs. *reject*) decision rule is associated with an arguably unnatural binary loss function (Gelman et al., 2013a). This rule also leads to an inability to ascertain simultaneous misfit (i.e., a lack of fit for both models under comparison) or to detect the presence of outliers.
- ✓ There is a lack of assessment of the uncertainty associated with the decision itself.
- ✓ The computation of marginal likelihoods is difficult in most settings (Chen et al., 2000; Marin and Robert, 2011) with further controversy about which solution to adopt (Newton and Raftery, 1994; Neal, 1994; Green, 1995; Chib, 1995; Neal, 1999; Skilling, 2006; Steele et al., 2006; Chopin and Robert, 2010).
- ✓ There is a strong dependence of the values of posterior probabilities on conditioning statistics, which in turn undermines their validity for model assessment. This is discussed in the setting of Approximate Bayesian computation (ABC) by Robert et al. (2011) and Marin et al. (2014).
- ✓ Finally, there is a temptation to create pseudo-frequentist equivalents such as q -values (Johnson and Rossell, 2010; Johnson, 2013a,b) with even less Bayesian justification.

Despite the many commentaries on these issues and attempts to address them (Berger and Jefferys, 1992; Madigan and Raftery, 1994; Balasubramanian, 1997; MacKay, 2002; Consonni et al., 2013; Jeffreys, 1939; Schwarz, 1978; Csizsár and Shields, 2000; Spiegelhalter et al., 2002; Plummer, 2008), there is still no agreed solution. We therefore propose a paradigm shift in

Bayesian hypothesis testing and model selection that resolves many of the above issues. In particular, the new perspective allows for the use of improper priors, provides a natural interpretation of the relative strengths of the models under consideration, delivers measures of uncertainty that can assist in decision-making, and is computationally feasible.

The proposed approach relies on the simple representation of the hypothesis test or model selection as a mixture estimation problem, where the weights are formally equal to 0 or 1. The mixture model (Frühwirth-Schnatter, 2006) thus contains the models under comparison as extreme cases. This approach is inspired from the consistency result of Rousseau and Mengersen (2011) on estimated overfitting mixtures, where the authors established that over-parameterised mixtures can be consistently estimated, despite the parameter lying on one or several boundaries of the parameter space.

The mixture representation is not directly equivalent to the traditional use of the posterior probability of a model given the data, i.e., the posterior estimator of the mixture weight is not a direct proxy for the posterior probability $\mathbb{P}(\mathfrak{M}_1|x)$. However, the posterior distribution of the mixture weights has the potential to be more informative about testing and model comparison. Moreover, the mixture approach has the additional appeal of not expanding the number of parameters in the model and hence is more aligned with Occam’s razor; see, for example, Adams (1987); Jefferys and Berger (1992); Rasmussen and Ghahramani (2001).

The plan of the paper is as follows: Section 2.1 provides a description of the mixture model specifically created for this setting, while Section 2.2 details the implementation issues with estimating the parameters of the mixture. Section 3 describes the performance of the mixture approach for standard i.i.d. models and Section 4 demonstrates its application to a case study that aims to identify the most appropriate model to describe survival of cancer patients. Section 5 expands on Rousseau and Mengersen (2011) to provide conditions on the hyperparameters of the mixture model that are sufficient to achieve convergence. Section 6 concludes on the generic applicability of the above principle.

2. TESTING PROBLEMS AS ESTIMATING MIXTURE MODELS

2.1 A new paradigm for testing

Given two competing statistical models,

$$\mathfrak{M}_1 : x \sim f_1(x|\theta_1), \theta_1 \in \Theta_1 \quad \text{and} \quad \mathfrak{M}_2 : x \sim f_2(x|\theta_2), \theta_2 \in \Theta_2,$$

which may correspond to a hypothesis to be tested and its alternative, respectively, it is always possible to embed both models within an encompassing mixture model

$$(1) \quad \mathfrak{M}_\alpha : x \sim \alpha f_1(x|\theta_1) + (1 - \alpha) f_2(x|\theta_2), \quad 0 \leq \alpha \leq 1.$$

Indeed, both models \mathfrak{M}_1 and \mathfrak{M}_2 correspond to very special cases of the mixture model, \mathfrak{M}_α , one for $\alpha = 1$ and the other for $\alpha = 0$ (with a slight notational inconsistency in the indices).

The choice of possible encompassing models is very general: for instance, a Geometric mixture

$$x \sim f_\alpha(x) \propto f_1(x|\theta_1)^\alpha f_2(x|\theta_2)^{1-\alpha}$$

is a conceivable alternative. However, some of these alternatives are less practical to manage. For example, for the Geometric mixture, the normalising constant is intractable, while, when f_1 and f_2 are Gaussian densities, the mixture remains Gaussian for all values of α . Similar drawbacks can be found with harmonic mixtures.

When considering a sample (x_1, \dots, x_n) from one of the two models, the mixture representation still holds at the likelihood level, namely the likelihood for each model is a special case

of the weighted sum of both likelihoods. However, this is not directly appealing for estimation purposes since it corresponds to a mixture with *a single observation*. See however O'Neill and Kypraios (2014) for a computational solution based upon this representation.

What we propose in this paper is to draw inference from the mixture representation (1) itself, acting as if each observation were individually and independently produced by the mixture model. An extension of the iid case is considered in Example 3.6 for linear models. Dependent observations like Markov chains can be modelled by a straightforward extension of (1) where both terms in the mixture are conditional on the relevant past observations.

Six principal advantages to this paradigm shift are as follows.

First, if the data were indeed generated from model \mathfrak{M}_1 , then relying on a Bayesian estimate of the weight α rather than on the posterior probability of model \mathfrak{M}_1 produces an equally convergent indicator of preference for this model (see Section 5).

Second, the mixture approach also removes the need for artificial prior probabilities on the model indices, ω_1 and ω_2 . Prior modelling only involves selecting an operational prior on α , for instance a Beta $\mathcal{B}(a_0, a_0)$ distribution, with a wide range of acceptable values for the hyperparameter a_0 , as demonstrated in Section 5. While the value of a_0 impacts on the posterior distribution of α , it can be argued that (a) it nonetheless leads to an accumulation of the mass near 1 or 0, and (b) a sensitivity analysis on the impact of a_0 is straightforward to carry out. Moreover, in most settings, this approach can be easily calibrated by a parametric bootstrap experiment providing a posterior distribution of α under each of the models under comparison. The prior predictive error can therefore be directly estimated and can drive the choice of the hyperparameter a_0 , if need be.

Third, the posterior distribution of α evaluates more thoroughly the strength of the support for a given model than the single figure outcome of a Bayes factor or of a posterior probability. The interpretation of α is at least as natural as the interpretation of the posterior probability, with the additional advantage of avoiding the zero-one loss setting (DeGroot, 1970, 1973; Berger, 1985). The quantity α and its posterior distribution provide a measure of proximity to both models for the data at hand, while being also interpretable as a propensity of the data to support, or stem from, one of the two models. The approach therefore allows for the possibility that, for a finite dataset, one model, both models or neither could be acceptable. This feature, which is missing from traditional Bayesian answers, will be seen in some of the illustrations below (Section 3).

Fourth, the problematic computation (Chen et al., 2000; Marin and Robert, 2011) of the marginal likelihoods is bypassed, since standard algorithms are available for Bayesian mixture estimation (Richardson and Green, 1997; Berkhof et al., 2003; Frühwirth-Schnatter, 2006; Lee et al., 2009). Another computational challenge that plagues Bayesian estimation for most mixture models is “label switching” (Celeux et al., 2000; Stephens, 2000; Jasra et al., 2005). This issue vanishes in this particular context, since components are no longer exchangeable. In particular, we compute neither a Bayes factor nor a posterior probability related to the substitute mixture model and we hence avoid the difficulty of recovering the modes of the posterior distribution (Berkhof et al., 2003; Lee et al., 2009; Rodriguez and Walker, 2014). Our perspective is solely centred on estimating the parameters of a mixture model where both components are always identifiable. Although using a Bayes factor to test for the number of components in the mixture (1) as in Richardson and Green (1997) would be possible, the outcome would fail to answer the original question of selecting between both (or more) models.

Fifth, the extension to a finite collection of models to be compared is straightforward, as this simply involves a larger number of components. This is discussed and illustrated in Section 4. The mixture approach allows consideration of all of these models at once, rather than engaging in costly pairwise comparisons. It is thus possible to eliminate the least likely models by simulation,

since these will not be explored by the corresponding computational algorithm [Carlin and Chib \(1995\)](#); [Richardson and Green \(1997\)](#). In addition, model averaging ([Hoeting et al., 1999](#); [Fernandez et al., 2001](#)) is automatically enhanced by this approach, with the further appeal of evaluating the allocation of each datapoint to one of the models under comparison.

The sixth and final advantage that we discuss here is that while standard (proper and informative) prior modelling can be painlessly reproduced in this novel setting, non-informative (improper) priors are also permitted, provided both models under comparison are first reparameterised so that they share common parameters. In the special case when all parameters can be made common to both models, the mixture model (1) can read as

$$\mathfrak{M}_\alpha : x \sim \alpha f_1(x|\theta) + (1 - \alpha)f_2(x|\theta), 0 \leq \alpha \leq 1.$$

For instance, if θ is a location parameter, a flat prior $\pi(\theta) \propto 1$ can be used with no foundational difficulty, in contrast to the testing case ([DeGroot, 1973](#); [Berger et al., 1998](#)).

We stress that while the requirement that the models share common parameters may seem extremely restrictive in a traditional mixture model, the presence of common parameters is quite natural within a testing setting. For instance, when comparing two different models for the *same* data, moments like $\mathbb{E}[X^\gamma]$ are defined in terms of the observed data and hence *should* be the *same* for both models. Reparametrising the models in terms of those common meaning moments does lead to a mixture model with some and maybe *all* common parameters. We thus advise the use of a common parameterisation, whenever possible.

Continuing from the previous argument, using at least some *identical* parameters on both components is a key differentiating feature of this reformulation of Bayesian testing, as it highlights the fact that the contrast between the two components of the mixture is *not* an issue of enjoying different parameters. In fact it is quite the opposite: the common parameters are nuisance parameters that need be integrated out, as in the traditional Bayesian approach through the computation of the marginal likelihoods. Note that even in the setting in which the parameters of the mixture components, θ_1 and θ_2 , differ, they can be integrated out by standard Monte Carlo methods.

We note that similar proposals have appeared in the econometrics literature, from the nesting approach of [Quandt \(1974\)](#), replacing a standard test with a less standard one ([Gouriéroux and Monfort, 1996](#)), to the geometric version of [Atkinson \(1970\)](#), with the issues of normalising constant and identification alluded to above, to the more general concept of encompassing ([Pesaran and Deaton, 1978](#)), or to the sequential prediction pool of [Geweke \(2010\)](#). However, to the extent of our knowledge, none of those approaches considered the embedded mixture from a Bayesian estimation perspective.

2.2 Mixture estimation

Before studying the application of the above principle to some standard examples in Section 3, we point out a few specifics of mixture estimation in such a particular setting. While the likelihood is a regular mixture likelihood, the fact that the weights are *a priori* close to the boundaries means that the usual completion approach of [Diebolt and Robert \(1994\)](#) is bound to be quite inefficient as soon as the sample size grows to moderate values. More precisely, if we consider a sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$ from (1) (or assumed to be from (1)), the completion of the sample by the latent component indicators ζ_i ($i = 1, \dots, n$) leads to the completed likelihood

$$(2) \quad \ell(\theta, \alpha_1, \alpha_2 \mid \mathbf{x}, \zeta) = \prod_{i=1}^n \alpha_{\zeta_i} f(x_i \mid \theta_{\zeta_i}) = \alpha^{n_1} (1 - \alpha)^{n_2} \prod_{i=1}^n f(x_i \mid \theta_{\zeta_i}),$$

where $(n_1, n_2) = (\sum_{i=1}^n \mathbb{I}_{\zeta_i=1}, \sum_{i=1}^n \mathbb{I}_{\zeta_i=2})$ under the constraint $n = \sum_{j=1}^2 \sum_{i=1}^n \mathbb{I}_{\zeta_i=j}$. This decomposition leads to a natural Gibbs implementation ([Diebolt and Robert, 1994](#)) where the

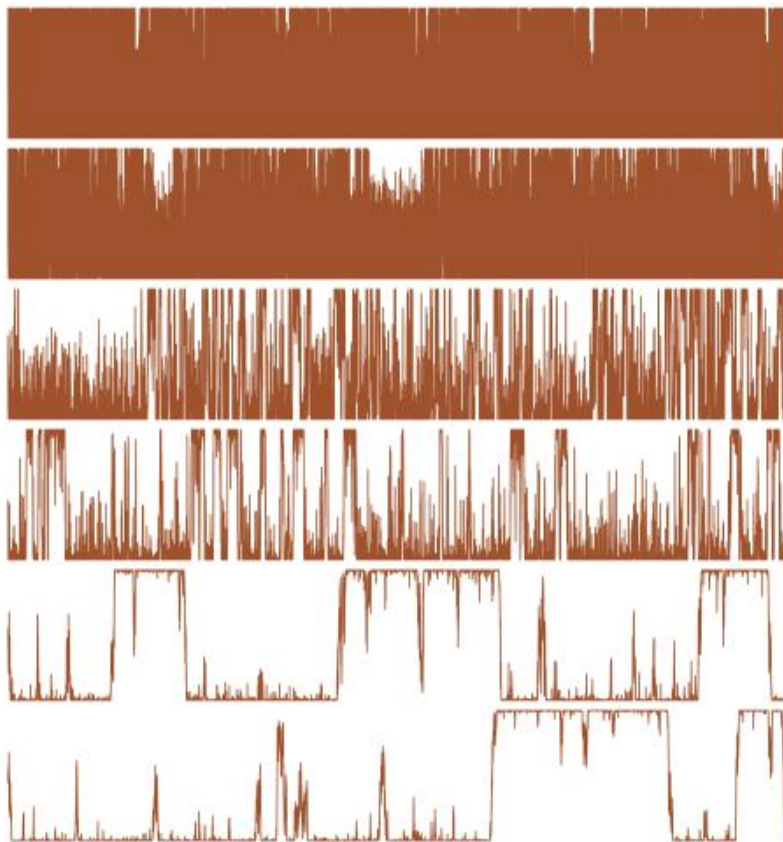


Fig 1: Gibbs sequences (α_t) on the first component weight for the mixture model $\alpha N(\mu, 1) + (1 - \alpha)N(0, 1)$ for a $N(0, 1)$ sample of size $N = 5, 10, 50, 100, 500, 10^3$ (from top to bottom) based on 10^5 simulations. The y -range range for all series is $(0, 1)$.

latent variables ζ_i and the parameters are generated from their respective conditional distributions. For instance, under a Beta $\mathcal{Be}(a_1, a_2)$ prior, α is generated from a Beta $\mathcal{Be}(a_1 + n_1, a_2 + n_2)$.

However, while this Gibbs sampling scheme is valid from a theoretical point of view, it faces convergence difficulties in the current setting, especially with large samples, due to the prior concentration on the boundaries of $(0, 1)$ for the mixture weight α . This feature is illustrated by Figure 1: as the sample size n grows, the Gibbs sample of the α 's exhibits much less switching between the vicinities of zero and one. The lack of label switching for regular mixture models is well-known and is due to the small probability of switching all component labels ζ_i at once; see, e.g., Celeux et al. (2000) and Lee et al. (2009). This issue is simply exacerbated in the current case due to extreme values for α .

Therefore, an alternative to the Gibbs sampler is needed (Lee et al., 2009) and we resort to a simple Metropolis-Hastings algorithm where the model parameters θ_i are generated from the respective posteriors of both models (that is, based on the entire sample) and where the mixture weight α is generated either from the prior distribution or from a random walk proposal on $(0, 1)$. It is indeed quite a rare occurrence for mixtures when we can use independent proposals. In the testing setting, the parameter θ_i can be considered independently within each model and its posterior can be based on the whole dataset. In cases when a common parameter is used in both components, one of the two available posteriors can be chosen at random at each iteration, either uniformly or based on the current value of α . The equivalent of Figure 1 for this Metropolis-Hastings implementation, Figure 2 exhibits a clear difference in the exploration

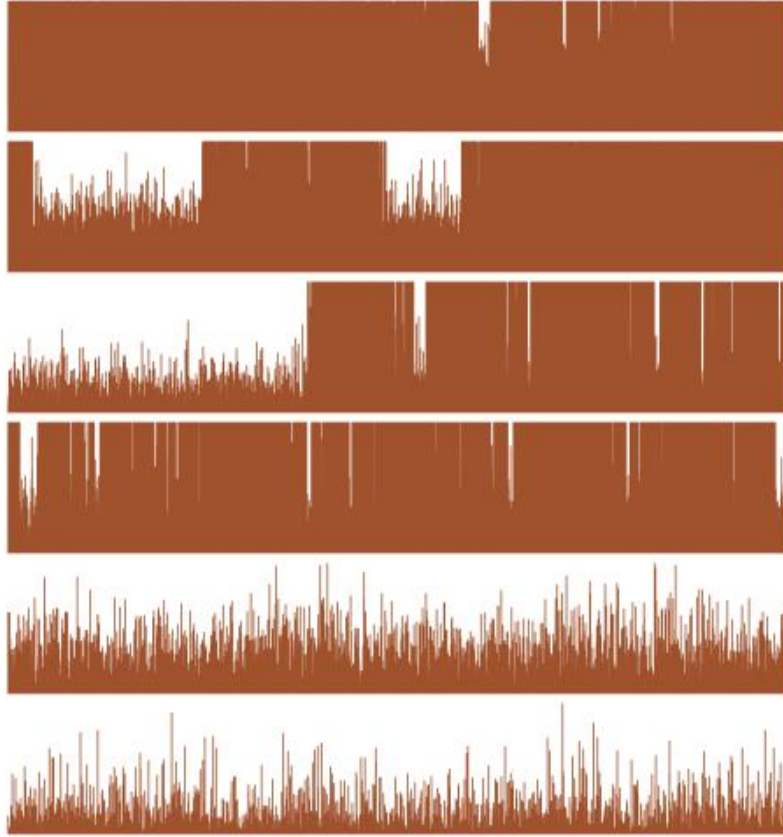


Fig 2: Metropolis-Hastings sequences (α_t) on the first component weight for the mixture model $\alpha N(\mu, 1) + (1 - \alpha)N(0, 1)$ for a $N(0, 1)$ sample of size $N = 5, 10, 50, 100, 500, 10^3$ (from top to bottom) based on 10^5 simulations. The y -range for all series is $(0, 1)$.

abilities of the resulting chain.

We also point out that the use of the posterior mean is highly inefficient in this situation. This is due to the specific pattern of the posterior distribution on α accumulating most of its weight on the endpoints of $(0, 1)$. We thus advocate that the posterior median be used instead as the relevant estimator of α .

3. ILLUSTRATIONS

In this Section, we proceed through a series of experiments in classical statistical settings in order to assess the ability of the mixture estimation approach to accurately differentiate between the models under comparison. As we will see throughout those examples, this experimentation also provides decisive confirmation of the consistency results obtained in Section 5. The first two examples are direct applications of Theorem 1 while the third is an application of Theorem 2.

Example 3.1 Consider a model choice test between a Poisson $\mathcal{P}(\lambda)$ distribution and a Geometric $\mathcal{Geo}(p)$ distribution, where the latter is defined as a number of failures and hence also starts at zero. We can represent the mixture (1) using the same parameter λ in the two distributions if we set $p = 1/(1+\lambda)$. The resulting mixture is then defined as

$$\mathfrak{M}_\alpha : \alpha \mathcal{P}(\lambda) + (1 - \alpha) \mathcal{Geo}(1/(1+\lambda))$$

This common parameterisation allows the use of Jeffreys' (1939) improper prior $\pi(\lambda) = 1/\lambda$, since the resulting posterior is then proper. Indeed, in a Gibbs sampling implementation, the full posterior distribution on λ , conditional on the allocation vector ζ , is given by

$$(3) \quad \pi(\lambda \mid \underline{x}, \zeta) \propto \exp(-n_1(\zeta)\lambda + \log\{\lambda\} (n\bar{x}_n - 1) (\lambda + 1)^{-\{n_2(\zeta) + s_2(\zeta)\}} ,$$

where $n_1(\zeta) = n - n_2(\zeta)$ is the number of observations allocated to the Poisson component, while $s_2(\zeta)$ is the sum of the observations that are allocated to the Geometric component. This conditional posterior is well defined for every ζ when $n > 0$, which implies that the marginal posterior is similarly well defined since ζ takes its values in a finite set. The distribution (3) can easily be simulated via an independent Metropolis-within-Gibbs step where the proposal distribution on λ is the Gamma distribution corresponding to the Poisson posterior. The motivation for this choice is that, since both distributions share the same mean parameter, using the posterior distribution associated with either one of the components and all the observations should be sufficiently realistic to produce high acceptance rates, even when the data are Geometric rather than Poisson. This strategy of relying on a model-based posterior as a proposal will be used throughout the examples. It obviously would not work for other mixture setups.

Under a $\mathcal{Be}(a_0, a_0)$ prior on α , the full conditional posterior density on α is a $\mathcal{Be}(n_1(\zeta) + a_0, n_2(\zeta) + a_0)$ distribution and the exact Bayes factor comparing the Poisson to the Geometric models is given by

$$\mathfrak{B}_{12} = n^{n\bar{x}_n} \prod_{i=1}^n x_i! \Gamma \left(n + 2 + \sum_{i=1}^n x_i \right) / \Gamma(n + 2).$$

However, this Bayes factor is undefined from a purely mathematical viewpoint, since it is associated with an improper prior on α (Jeffreys, 1939; DeGroot, 1973; Berger et al., 1998; Robert et al., 2009). The posterior probability of the Poisson model is derived as

$$\mathbb{P}(\mathfrak{M}_1 | x) = \frac{\mathfrak{B}_{12}}{1 + \mathfrak{B}_{12}}$$

when adopting (without much of a justification) identical prior weights on both models.

A first experiment in assessing our approach is based on 100 datasets simulated from a Poisson $\mathcal{P}(4)$ distribution. As shown in Figure 3, not only is the parameter λ properly estimated, but the estimation of α is very close to 1 for a sample size equal to $n = 1000$. In this case, the smaller the value of a_0 , the better in terms of proximity to 1 of the posterior distribution on α . Note that the choice of a_0 does not significantly impact on the posterior distribution of λ . Figure 4 gives an assessment of the convergence of the Metropolis-Hastings for λ and the mixture model weight α even if the sample size is very small ($n=5$).

Figure 5 demonstrates the convergence of the posterior means and posterior medians of α as the sample sizes n increase for the same Poisson $\mathcal{P}(4)$ simulated samples. The sensitivity of the posterior distribution of α on the hyperparameter a_0 is clearly depicted in the graph. While all posterior means and medians converge to 1 in this simulation, the impact of small values of a_0 on the estimates is such that we consider values $a_0 \leq .1$ as having too strong and too lengthy an influence on the posterior distribution to be acceptable.

Notwithstanding the fact that the analysis is invalid because of the reliance on improper priors, we also compare the outcome of a traditional Bayesian analysis with our estimates of α . Figure 6 shows how the posterior probability of model \mathfrak{M}_1 and the posterior median of α relate as the sample size grows to 1000. The shaded areas indicate the range of all estimates of α , which varies between .2 and .8 for $a_0 = .5$ and between 0 and 1 for $a_0 \leq .1$. This difference reinforces our earlier recommendation that small values of a_0 should be avoided, as they overwhelm the information contained in the data for small sample sizes.

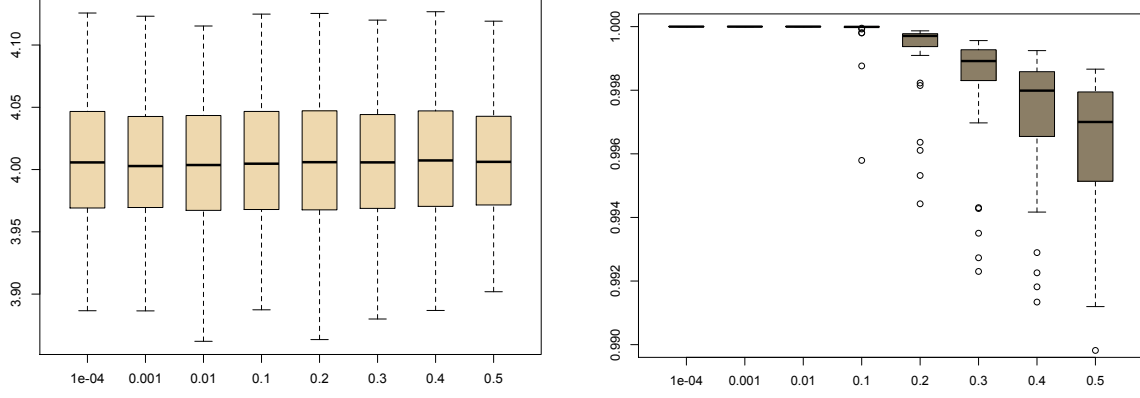


Fig 3: **Example 3.1:** Boxplots of the posterior means (*wheat*) of λ and the posterior medians (*dark wheat*) of α for 100 Poisson $\mathcal{P}(4)$ datasets of size $n = 1000$ for $a_0 = .0001, .001, .01, .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

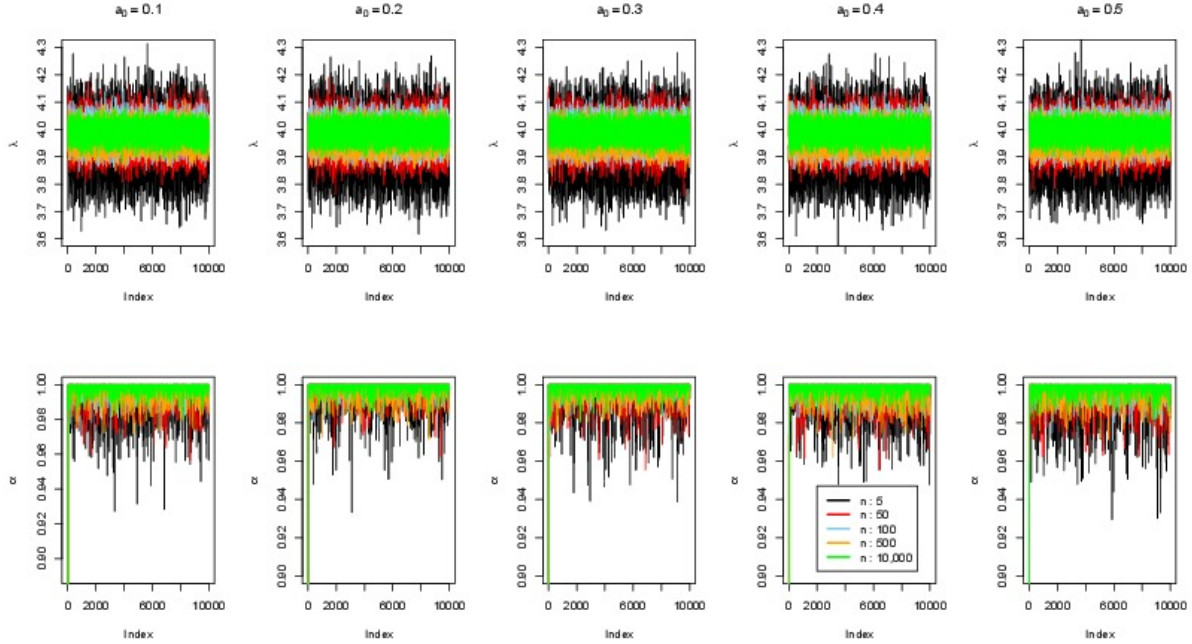


Fig 4: **Example 3.1:** Dataset from a Poisson distribution $\mathcal{P}(4)$: Estimations of (*Top*) λ and (*Bottom*) α via Metropolis-Hastings algorithm over 10^4 iterations for 5 samples of size $n = 5, 50, 100, 500, 10,000$.

A symmetric experiment is to study the behaviour of the posterior distribution on α for data from the alternative model, i.e., a Geometric distribution. Based on 100 datasets from a Geometric $\mathcal{G}(0.1)$ distribution, Figure 7 displays the very fast convergence of the posterior median to 0 for all values of a_0 considered, even though the impact of this hyperprior is noticeable.

Example 3.2 Consider a hypothesis test between a normal $\mathcal{N}(\theta_1, 1)$ and a normal $\mathcal{N}(\theta_2, 2)$ distribution. We again construct the mixture so that the same location parameter θ is used

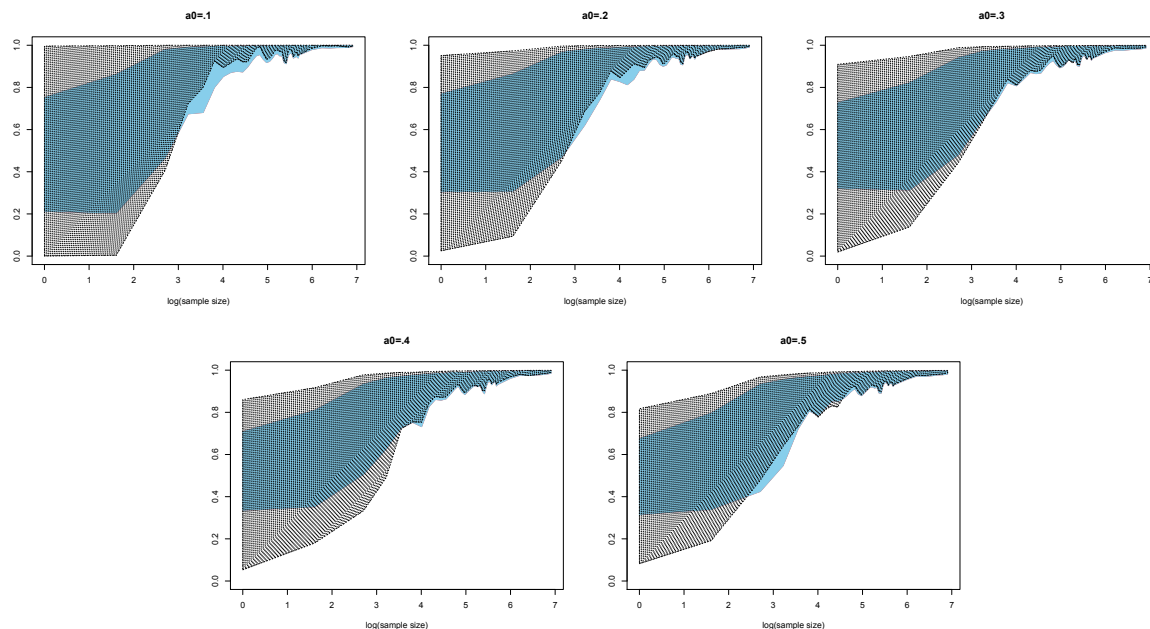


Fig 5: **Example 3.1:** Posterior means (*sky-blue*) and medians (*grey-dotted*) of the posterior distributions on α , displayed over 100 Poisson $\mathcal{P}(4)$ datasets for sample sizes from 1 to 1000. The shaded and dotted areas indicate the range of the estimates. Each plot corresponds to a Beta prior on α with parameter $a_0 = .1, .2, .3, .4, .5$ and each posterior approximation is based on 10^4 iterations.

in both distributions, which allows the use of Jeffreys' (1939) non-informative prior $\pi(\theta) = 1$. We thus embed the test in a mixture of normal models, $\alpha\mathcal{N}(\theta, 1) + (1 - \alpha)\mathcal{N}(\theta, 2)$, and adopt a Beta $\mathcal{B}(a_0, a_0)$ prior on α . In this case, the considering the posterior distribution on (α, θ) , conditional on the allocation vector ζ , displays conditional independence between θ and α :

$$\theta|\mathbf{x}, \zeta \sim \mathcal{N}\left(\frac{n_1\bar{x}_1 + .5n_2\bar{x}_2}{n_1 + .5n_2}, \frac{1}{n_1 + .5n_2}\right), \quad \alpha|\zeta \sim \mathcal{Be}(a_0 + n_1, a_0 + n_2),$$

where n_i and \bar{x}_i denote the number of observations and the empirical mean of the observations allocated to component i , respectively (with the convention that $n_i\bar{x}_i = 0$ when $n_i = 0$). Since this conditional posterior distribution is well-defined for every possible value of ζ and since the distribution ζ has a finite support, $\pi(\theta|x)$ is proper.

Note that for this example, the conditional evidence $\pi(x|\zeta)$ can easily be derived in closed form, which means that a random walk on the allocation space $\{1, 2\}^n$ could be implemented. We did not follow this direction, as it seemed unlikely such a random walk would have been more efficient than a Metropolis–Hastings algorithm on the parameter space only.

As in the previous example, in order to evaluate the convergence of the estimates of the mixture weights, we simulated 100 $\mathcal{N}(0, 1)$ datasets. Figure 8 displays the range of the posterior means and medians of α when either a_0 or n varies. The figure shows the same concentration effect (if a lingering impact of a_0) when n increases. We also included the posterior probability of \mathfrak{M}_1 in the comparison, derived from the Bayes factor

$$\mathfrak{B}_{12} = 2^{n-1/2} / \exp^{1/4} \sum_{i=1}^n (x_i - \bar{x})^2,$$

with equal prior weights, even though it is not formally well defined since it is based on an improper prior. The shrinkage of the posterior expectations towards 0.5 confirms our recom-

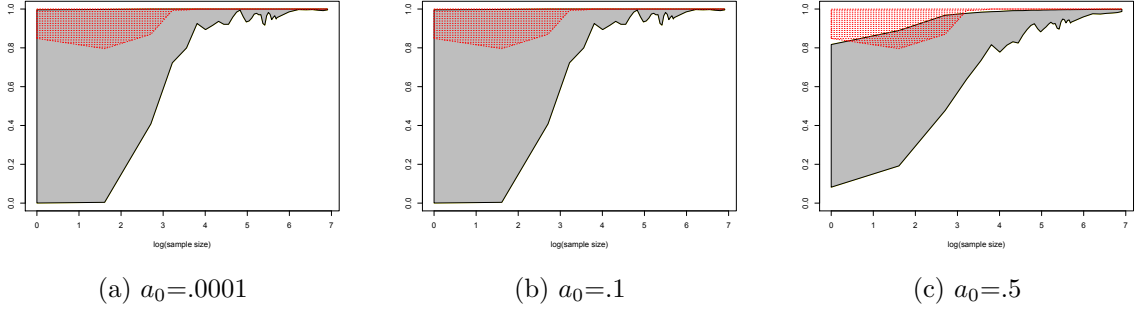


Fig 6: **Example 3.1:** Comparison between the ranges of $\mathbb{P}(\mathfrak{M}_1|x)$ (red dotted area) and of the posterior medians of α for 100 Poisson $\mathcal{P}(4)$ datasets with sample sizes n ranging from 1 to 1000 and for several values of the hyperparameter a_0 .

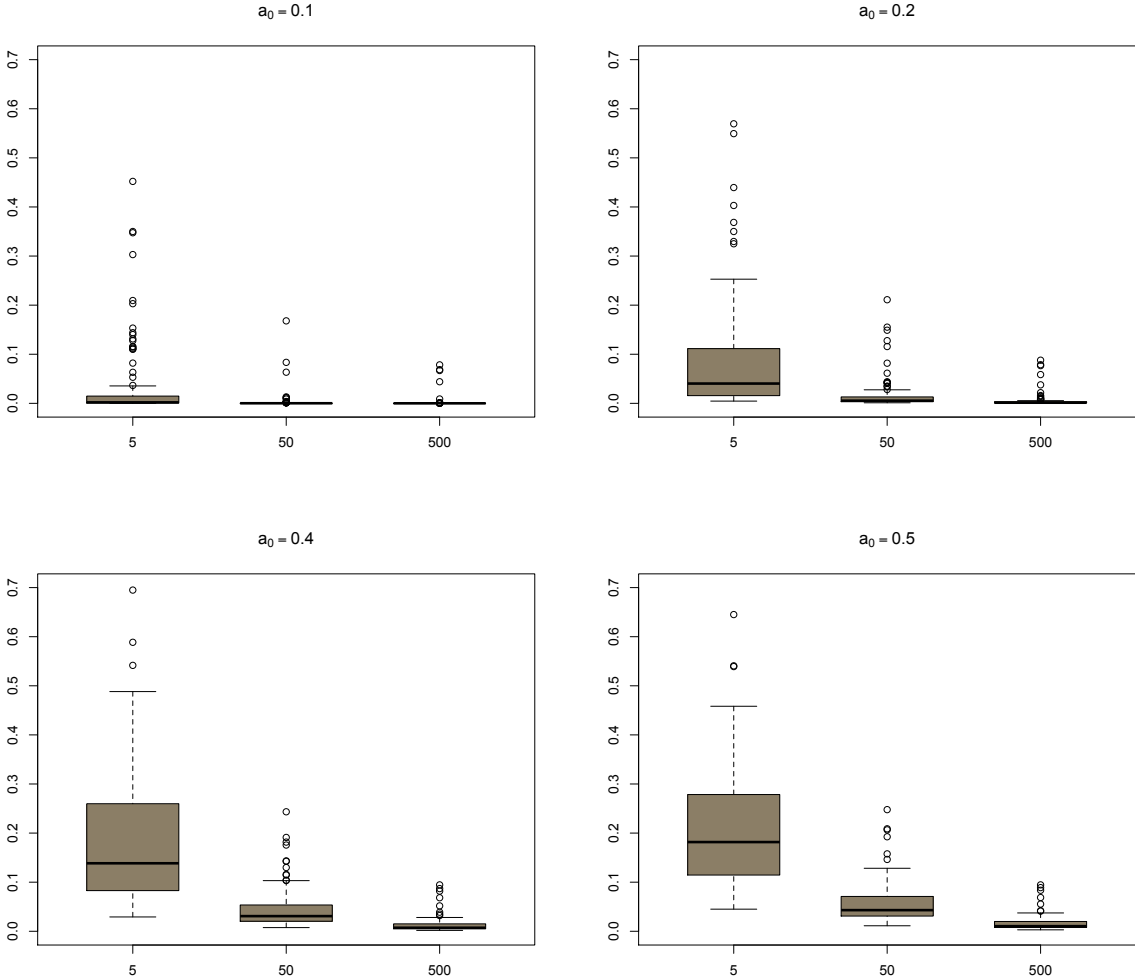


Fig 7: **Example 3.1:** Boxplots of the posterior medians of α for 100 geometric $\mathcal{Geo}(.1)$ datasets of size $n = 5, 50, 500$. Boxplots are plotted using four beta priors for α with $a_0 = .1, .2, .4, .5$. Each posterior approximation is based on 10^4 iterations.

mendation to use the posterior median instead. The same concentration phenomenon occurs for the $\mathcal{N}(0, 2)$ case, as illustrated in Figure 10 for a single $\mathcal{N}(0, 2)$ dataset.

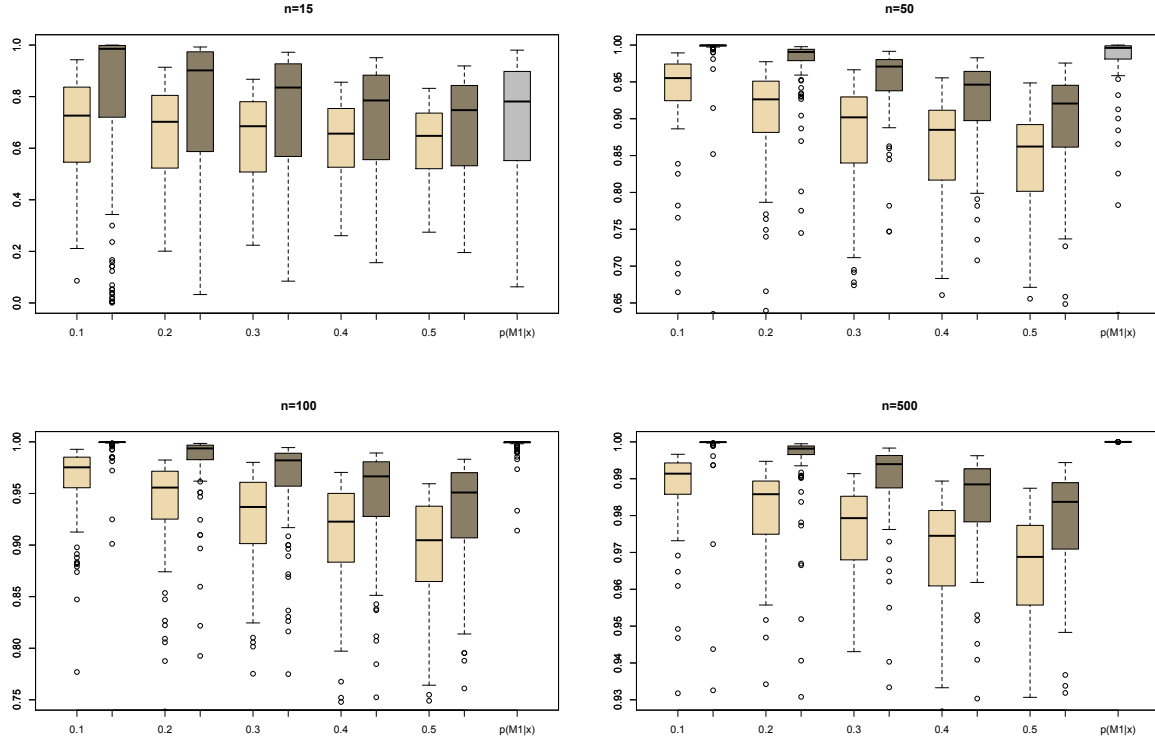


Fig 8: **Example 3.2:** Boxplots of the posterior means (*wheat*) and medians of α (*dark wheat*), compared with a boxplot of the exact posterior probabilities of \mathfrak{M}_0 (*gray*) for a $\mathcal{N}(0, 1)$ sample, derived from 100 datasets for sample sizes equal to 15, 50, 100, 500. Each posterior approximation is based on 10^4 MCMC iterations.

In order to better understand the nature of the convergence of the posterior distribution of α to the proper limiting value, we present in Figure 9 a magnified version of the behaviour of $\log(n) \log(1 - \mathbb{E}[\alpha|\mathbf{x}])$ and $\log(1 - p(\mathfrak{M}_1|\mathbf{x}))$ as the sample size n grows. Most interestingly, the variation is similar for both procedures, even though the choice of the hyperparameter a_0 impacts on the variability of the mixture solution. This is due to the fact that the asymptotic regime is not quite reached for these sample sizes, as $1 - \mathbb{P}(\mathfrak{M}_1|\mathbf{x}) \leq e^{-cn}$ for some positive c with high probability, while $\mathbb{E}[\alpha|\mathbf{x}] = O(n^{-1/2})$, leading to

$$\log(n) \log(1 - \mathbb{E}[\alpha|\mathbf{x}]) \asymp -(\log n)^2.$$

Furthermore, the alternative of considering the posterior probability of having the *entire sample* being generated from a single component is not relevant for the comparison as this estimate is almost always very close to zero. This means that while α captures the model preferred by the data, in only extreme cases would the mixture model favour a completely homogeneous sample that would come from one and only one component. In comparison, if we had instead called the algorithm of [van Havre et al. \(2014\)](#), we would have obtained mostly homogeneous samples for very small values of a_0 . Their algorithm is a special type of tempering MCMC, where tempering is obtained by choosing successive values of a_0 , ranging from large to very small.

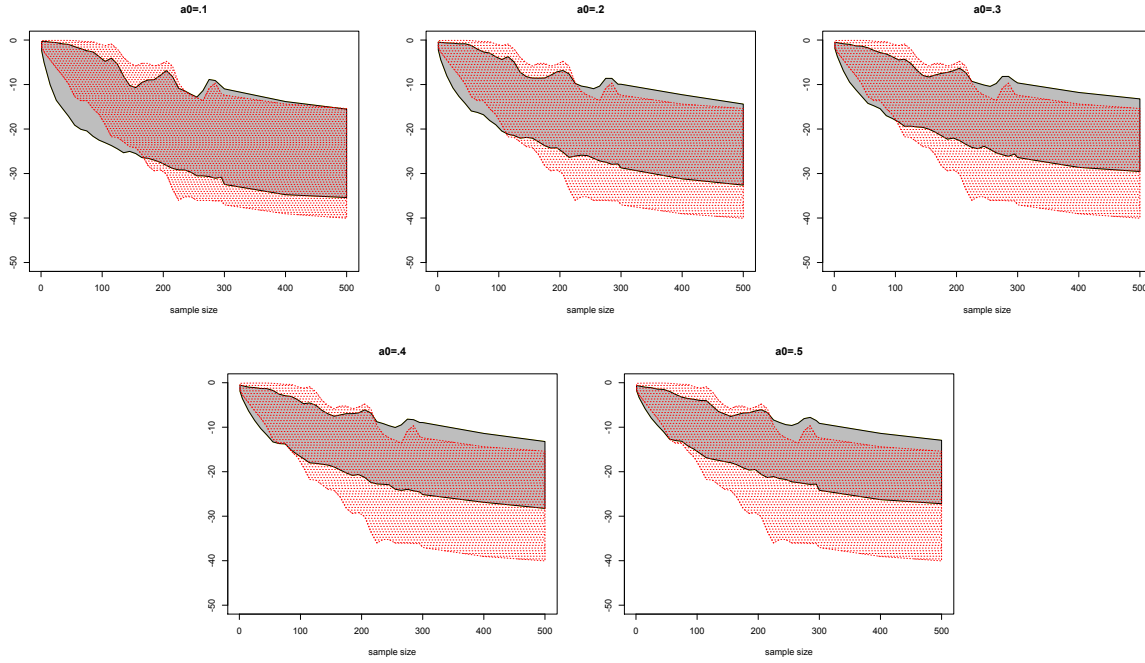


Fig 9: **Example 3.2:** Plots of ranges of $\log(n) \log(1 - \mathbb{E}[\alpha|x])$ (gray colour) and $\log(1 - p(\mathcal{M}_1|x))$ (red dotted) over 100 $\mathcal{N}(0, 1)$ samples as sample size n grows from 1 to 500. and α is the weight of $\mathcal{N}(0, 1)$ in the mixture model. The shaded areas indicate the range of the estimations and each plot is based on a Beta prior with $a_0 = .1, .2, .3, .4, .5, 1$ and each posterior approximation is based on 10^4 iterations.

Example 3.3 We now consider a setting in which we oppose a $\mathcal{N}(0, 1)$ model against a $\mathcal{N}(\mu, 1)$ model, hence testing whether or not $\mu = 0$. Since this is an embedded case, we cannot use an improper prior on μ and thus settle for a $\mu \sim \mathcal{N}(0, 1)$ prior. As discussed above in Section 2.2, Gibbs sampling applied to this mixture posterior model shows poor performance and should be replaced with a Metropolis–Hastings algorithm.

The resulting inference on the weight of the $\mathcal{N}(\mu, 1)$ component, α , contrasts the possibility that the data are distributed as $\mathcal{N}(0, 1)$ with the possibility that they are not from this null distribution. In the former case, obtaining values of α close to one requires larger sample sizes than in the latter case. Figure 11 displays the behaviour of the posterior distribution of α when the sample comes from a normal distribution $\mathcal{N}(1, 1)$. For a sample of size 10^2 , the accumulation of α on $(.8, 1)$ illustrates the strength of the support for the model $\mathcal{N}(\mu, 1)$. This support reduces as a_0 increases. The impact of the small sample size on the posterior distributions of α is shown in the right hand side of Figure 11 for the case in which $a_0 = .1$; it is apparent that for $n = 5$ we cannot assess which model best fits the data.

Example 3.4 Inspired from Marin et al. (2014), we oppose the normal $\mathcal{N}(\mu, 1)$ model to the double-exponential $\mathcal{L}(\mu, \sqrt{2})$ model. The scale $\sqrt{2}$ is intentionally chosen to make both distributions share the same variance. As in the normal case in Example 3.2, the location parameter μ can be shared by both models and allows for the use of the flat Jeffreys’ prior. As in all previous examples, Beta distributions $\mathcal{B}(a_0, a_0)$ are compared wrt their hyperparameter a_0 .

Whereas in the previous examples we illustrated that the posterior distribution of the weight of the true model converged to 1, we now consider a setting in which neither model is correct. We achieve this by using a $\mathcal{N}(0, .7^2)$ to simulate the data. In this specific case, both posterior

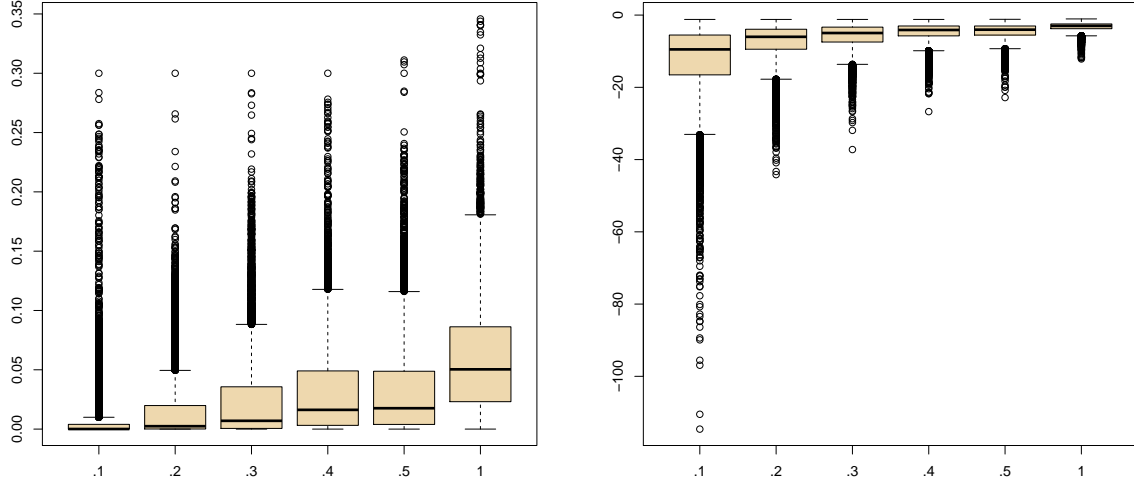


Fig 10: **Example 3.2:** (left) Posterior distributions of the mixture weight α and (right) of their logarithmic transform $\log\{\alpha\}$ under a Beta $\mathcal{B}(a_0, a_0)$ prior when $a_0 = .1, .2, .3, .4, .5, 1$ and for a normal $\mathcal{N}(0, 2)$ sample of 10^3 observations. The MCMC outcome is based on 10^4 iterations.

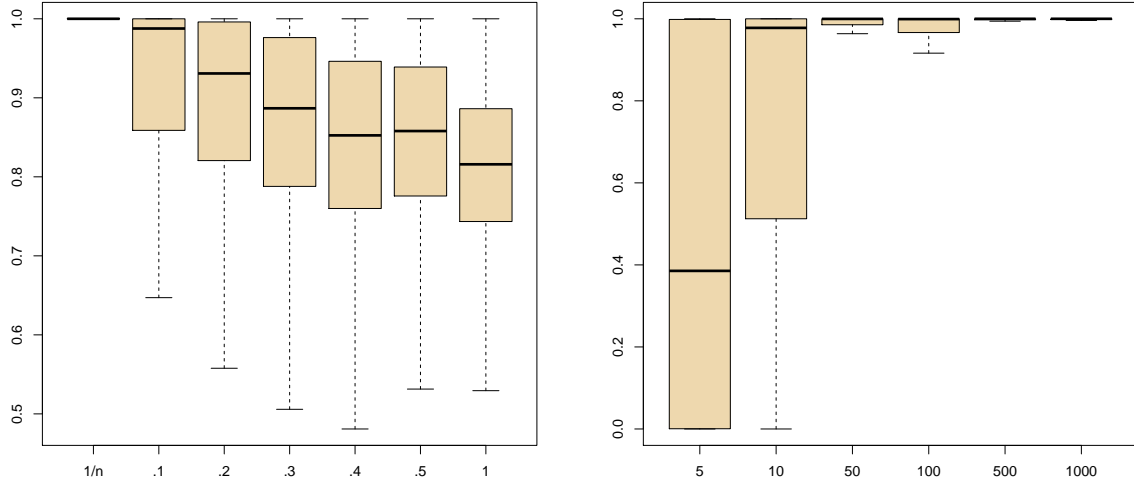


Fig 11: **Example 3.3** Posterior distributions of the $\mathcal{N}(\mu, 1)$ component weight α under a Beta $\mathcal{B}(a_0, a_0)$ prior (left) for $a_0 = 1/n, .1, .2, .3, .4, .5, 1$ with 10^2 $\mathcal{N}(1, 1)$ observations and (right) for $a_0 = .1$ with $n = 5, 10, 50, 100, 500, 1000$ $\mathcal{N}(1, 1)$ observations. In both cases each posterior approximation is based on 10^5 MCMC iterations.

means and medians of α fail to concentrate near 0 and 1 as the sample size increases, as shown on Figure 12. Thus in a majority of cases in this experiment, the outcome indicates that neither of both models is favoured by the data. This example does not exactly follow the assumptions of Theorem 1 since the Laplace distribution is not differentiable everywhere. However, it is almost surely differentiable and it is differentiable in quadratic mean, so we expect to see the same types of behaviour as predicted by Theorem 1.

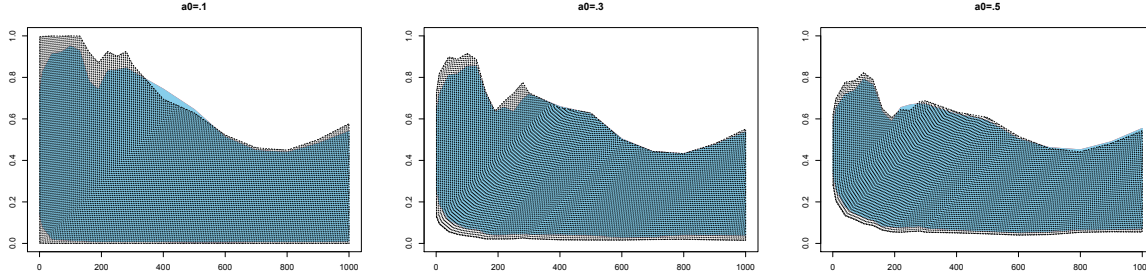


Fig 12: **Example 3.4:** Ranges of posterior means (*skyblue*) and medians (*dotted*) of the weight α of model $\mathcal{N}(\theta, 1)$ over 100 $\mathcal{N}(0, .7^2)$ datasets for sample sizes from 1 to 1000. Each estimate is based on a Beta prior with $a_0 = .1, .3, .5$ and 10^4 MCMC iterations.

In this example, the Bayes factor associated with Jeffreys' prior is defined as

$$\mathfrak{B}_{12} = \frac{\exp \left\{ -\sum_{i=1}^n (x_i - \bar{x})^2 / 2 \right\}}{(\sqrt{2\pi})^{n-1} \sqrt{n}} \bigg/ \int_{-\infty}^{\infty} \frac{\exp \left\{ -\sum_{i=1}^n |x_i - \mu| / \sqrt{2} \right\}}{(2\sqrt{2})^n} d\mu$$

where the denominator is available in closed form (see Appendix 1). As above, since the prior is improper, this quantity is formally undefined. Nonetheless, we employ it in order to compare Bayes estimators of α with the posterior probability of the model being a $\mathcal{N}(\mu, 1)$ distribution. Based on a Monte Carlo experiment involving 100 replicates of a $\mathcal{N}(0, .7^2)$ dataset, Figure 13 demonstrates the reluctance of the estimates of α to approach 0 or 1, while $\mathbb{P}(\mathfrak{M}_1 | \mathbf{x})$ varies over the whole range of 0 and 1 for all sample sizes considered here. While this is a weakly informative indication, the right hand side of Figure 13 shows that, on average, the posterior estimates of α converge toward a value between .1 and .4 for all a_0 while the posterior probabilities converge to .6. In that respect, both criteria offer a similar interpretation about the data because neither α nor $P(\mathfrak{M}_1 | x)$ indicate strong support for either model.

In the following two examples we consider regression models. The theory and methodology established above can be extended to this case under the assumption that the design is random. Hence example 3.5 is an application of Theorem 1 while example 3.6 is an application of Theorem 2.

Example 3.5 In this example, we apply our testing strategy to a binary response, using the R dataset about diabetes in Pima Indian women (R Development Core Team, 2006) as a benchmark (Marin and Robert, 2007). This dataset pertains to a randomly selected sample of 200 women tested for diabetes according to WHO criteria. The response variable y is “Yes” or “No”, for presence or absence of diabetes and the explanatory variable \mathbf{x} is restricted here to the bmi , body mass index weight in $\text{kg}/(\text{height in m})^2$. For this problem, either logistic or probit regression models could be suitable. We are thus comparing both fits via our method. If $\mathbf{y} = (y_1 \ y_2 \dots y_n)$ is the vector of binary responses and $X = [I_n \ \mathbf{x}_1]$ is the $n \times 2$ matrix of

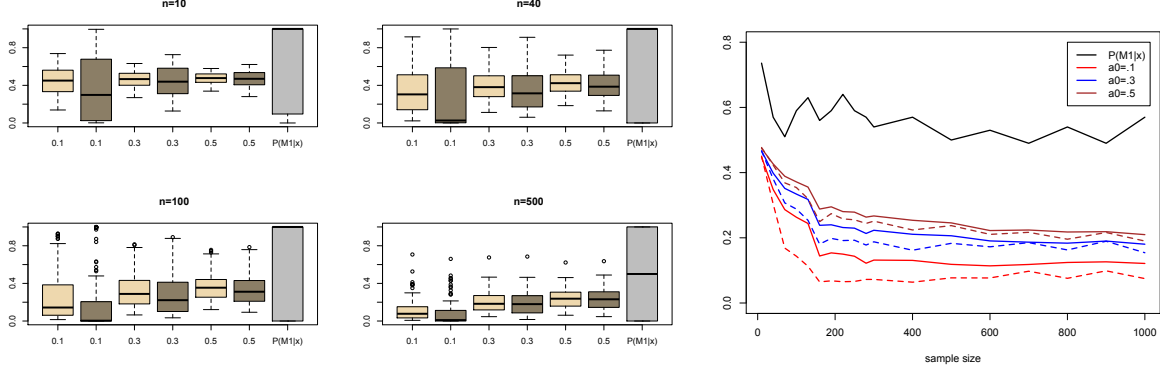


Fig 13: **Example 3.4:** (left) Boxplot of the posterior means (*wheat*) and medians (*dark wheat*) of α , and of the posterior probabilities of model $\mathcal{N}(\mu, 1)$ over 100 $\mathcal{N}(0, .7^2)$ datasets for sample sizes $n = 10, 40, 100, 500$; (right) averages of the posterior means and posterior medians of α against the posterior probabilities $\mathbb{P}(\mathfrak{M}_1|\mathbf{x})$ for sample sizes going from 1 to 1000. Each posterior approximation is based on 10^4 Metropolis-Hastings iterations.

corresponding explanatory variables, the models in competition can be defined as ($i = 1, \dots, n$)

$$(4) \quad \begin{aligned} \mathfrak{M}_1 : y_i | \mathbf{x}^i, \theta_1 &\sim \mathcal{B}(1, p_i) \quad \text{where} \quad p_i = \frac{\exp(\mathbf{x}^i \theta_1)}{1 + \exp(\mathbf{x}^i \theta_1)} \\ \mathfrak{M}_2 : y_i | \mathbf{x}^i, \theta_2 &\sim \mathcal{B}(1, q_i) \quad \text{where} \quad q_i = \Phi(\mathbf{x}^i \theta_2) \end{aligned}$$

where $\mathbf{x}^i = (1 \ x_{i1})$ is the vector of explanatory variables corresponding to the i th response and $\theta_j = (\theta_{0j}, \theta_{1j})$, $j = 1, 2$, is a 2×1 vector comprising the intercept and regression coefficient, respectively under either \mathfrak{M}_1 or \mathfrak{M}_2 . We once again consider the case in which both models share the same parameter.

For this generalised linear model, there is no moment equation that relates θ_1 and θ_2 . We thus adopt a local reparameterisation strategy by rescaling the parameters of the probit model \mathfrak{M}_2 so that the MLE's of both models coincide. This strategy follows from the remark by [Choudhury et al. \(2007\)](#) about the connection between the normal cdf and a logistic function

$$\Phi(\mathbf{x}^i \theta_2) \approx \frac{\exp(k \mathbf{x}^i \theta_2)}{1 + \exp(k \mathbf{x}^i \theta_2)}$$

and we attempt to find the best estimate of k to bring both sets of parameters into coherency. Given

$$(k_0, k_1) = (\widehat{\theta}_{01}/\widehat{\theta}_{02}, \widehat{\theta}_{11}/\widehat{\theta}_{12}),$$

which denote ratios of the maximum likelihood estimates of the logistic model parameters to those for the probit model, we redefine q_i in (4) as

$$(5) \quad q_i = \Phi(\mathbf{x}^i (\kappa^{-1} \theta))$$

$$\kappa^{-1} \theta = (\theta_0/k_0, \theta_1/k_1).$$

Once the mixture model is thus parameterised, we set our now standard Beta $\mathcal{B}(a_0, a_0)$ on the weight of \mathfrak{M}_1 , α , and choose the default g -prior on the regression parameter (see, e.g., [Marin and Robert, 2007](#), Chapter 4), so that

$$\theta \sim \mathcal{N}_2(0, n(X^T X)^{-1}).$$

TABLE 1
Dataset *Pima.tr*: Posterior medians of the mixture model parameters.

| a_0 | Logistic model parameters | | | | Probit model parameters | |
|-------|---------------------------|------------|------------|--|-------------------------|------------------------|
| | α | θ_0 | θ_1 | | $\frac{\theta_0}{k_0}$ | $\frac{\theta_1}{k_1}$ |
| .1 | .352 | -4.06 | .103 | | -2.51 | .064 |
| .2 | .427 | -4.03 | .103 | | -2.49 | .064 |
| .3 | .440 | -4.02 | .102 | | -2.49 | .063 |
| .4 | .456 | -4.01 | .102 | | -2.48 | .063 |
| .5 | .449 | -4.05 | .103 | | -2.51 | .064 |

TABLE 2
Simulated dataset: Posterior medians of the mixture model parameters.

| True model: | \mathfrak{M}_α^1 logistic with $\theta_1 = (5, 1.5)$ | | | | | \mathfrak{M}_α^2 probit with $\theta_2 = (3.5, .8)$ | | | | |
|-------------|--|------------|------------|------------------------|------------------------|---|------------|------------|------------------------|------------------------|
| | α | θ_0 | θ_1 | $\frac{\theta_0}{k_0}$ | $\frac{\theta_1}{k_1}$ | α | θ_0 | θ_1 | $\frac{\theta_0}{k_0}$ | $\frac{\theta_1}{k_1}$ |
| .1 | .998 | 4.940 | 1.480 | 2.460 | .640 | .003 | 7.617 | 1.777 | 3.547 | .786 |
| .2 | .972 | 4.935 | 1.490 | 2.459 | .650 | .039 | 7.606 | 1.778 | 3.542 | .787 |
| .3 | .918 | 4.942 | 1.484 | 2.463 | .646 | .088 | 7.624 | 1.781 | 3.550 | .788 |
| .4 | .872 | 4.945 | 1.485 | 2.464 | .646 | .141 | 7.616 | 1.791 | 3.547 | .792 |
| .5 | .836 | 4.947 | 1.489 | 2.465 | .648 | .186 | 7.596 | 1.782 | 3.537 | .788 |

In a Gibbs representation (not implemented here), the full conditional posterior distributions given the allocation vector ζ are $\alpha \sim \mathcal{B}(a_0 + n_1, a_0 + n_2)$ and

$$(6) \quad \pi(\theta \mid \mathbf{y}, X, \zeta) \propto \frac{\exp\{\sum_i \mathbb{I}_{\zeta_i=1} y_i \mathbf{x}^i \theta\}}{\prod_{i; \zeta_i=1} [1 + \exp(\mathbf{x}^i \theta)]} \exp\{-\theta^T (X^T X) \theta / 2n\} \\ \times \prod_{i; \zeta_i=2} \Phi(\mathbf{x}^i (\kappa^{-1} \theta))^{y_i} (1 - \Phi(\mathbf{x}^i (\kappa^{-1} \theta)))^{(1-y_i)}$$

where n_1 and n_2 are the number of observations allocated to the logistic and probit models, respectively. This conditional representation shows that the posterior distribution is then clearly defined, which is obvious when considering that for once the chosen prior is proper.

For the Pima dataset, the maximum likelihood estimates of the regression parameters are $\hat{\theta}_1 = (-4.11, 0.10)$ and $\hat{\theta}_2 = (-2.54, 0.065)$, respectively, so $k = (1.616, 1.617)$. We compare the outcomes of this Bayesian analysis when $a_0 = .1, .2, .3, .4, .5$ in Table 1. As clearly shown by the Table, the estimates of α are close to 0.5 for all values of a_0 , and the estimates of θ_0 and θ_1 are very stable (and quite similar to the MLEs). We note a slight inclination of α towards 0.5 as a_0 increases, but do not want to over-interpret the phenomenon. This behaviour leads us to conclude that (a) neither or both of the models are appropriate for the Pima Indian data; (b) the sample size may be too small to allow discrimination between the logit and the probit models.

To follow up on this last remark, we ran a second experiment with simulated logit and probit datasets and a larger sample size $n = 10,000$. We used the regression coefficients $(5, 1.5)$ for the logit model, and $(3.5, .8)$ for the probit model. The estimates of the parameters of both \mathfrak{M}_{α_1} and \mathfrak{M}_{α_2} and for both datasets are produced in Table 2. For every a_0 , the estimates in the true model are quite close to the true values and the posterior estimates of α are either close to 1 in the logit case and to 0 in the probit case. For this large setting, there is thus consistency in the selection of the proper model. In addition, Figure 14 shows that when the sample size is large enough, the posterior distribution of α concentrates its mass near 1 and 0 when the data are simulated from a logit and a probit model, respectively.

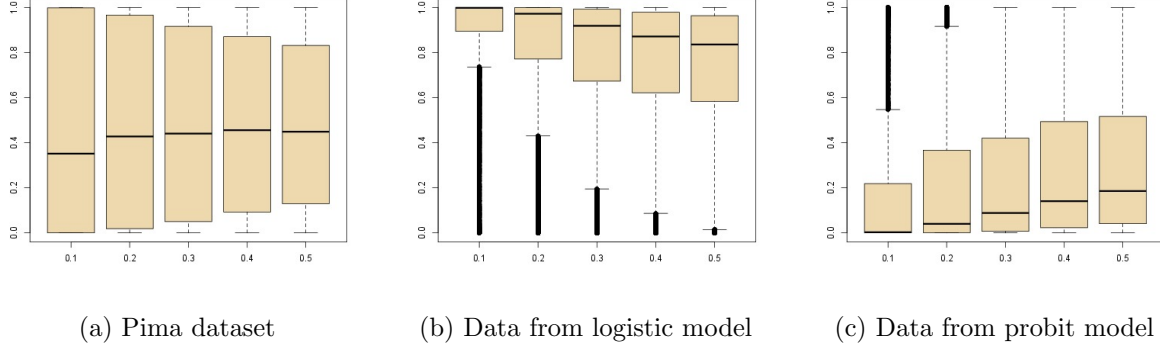


Fig 14: **Example 3.5:** Histograms of the posterior distributions of α in favour of the logistic model based on 10^4 Metropolis-Hastings iterations where $a_0 = .1, .2, .3, .4, .5$.

Example 3.6 We turn now to the classical issue of variable selection in a Gaussian linear regression model. Given a vector of outcomes (y_1, y_2, \dots, y_n) and the corresponding explanatory variables represented by the $n \times (k+1)$ matrix $X = [\mathbf{1}_n \ X_1 \ \dots \ X_k]$ (including $\mathbf{1}_n$, a first column of 1's), we assume that

$$(7) \quad y \mid X, \beta, \sigma^2 \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$$

where $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$ is a $k+1$ -vector of $k+1$ elements with β_0 the intercept. If we consider the generic case in which any covariate could be removed from the model, we are facing the comparison of $2^{k+1} - 1$ models, corresponding to every possible subset of explanatory variables. In our framework, this means evaluating a mixture model (1) with $\gamma = 2^{k+1} - 1$ components. For $j = 1, \dots, \gamma$, \mathfrak{M}_j will denote the corresponding model, v_j the number of explanatory variables used in \mathfrak{M}_j , β^j the vector of the v_j regression coefficients and X^j the sub-matrix of X derived from the covariate variables included in \mathfrak{M}_j .

The corresponding mixture model used for testing is therefore given by

$$(8) \quad \mathfrak{M}_\alpha : y \sim \sum_{j=1}^{\gamma} \alpha_j \mathcal{N}(X^j \beta^j, \sigma^2 I_n) \quad \sum_{j=1}^{\gamma} \alpha_j = 1.$$

When introducing a missing variable representation, each observation y_i is associated with a missing variable ζ_i taking values in $1, 2, \dots, \gamma$. The weights of the mixture (8) are associated with a symmetric Dirichlet prior $(\alpha_1, \dots, \alpha_\gamma) \sim \mathcal{D}_\gamma(a_0, \dots, a_0)$.

Contrary to the previous examples of this section, we now consider two different settings, corresponding to the separate versus common parameterisations of the different models \mathfrak{M}_j .

Case 1. If \mathfrak{M}_f denotes the full regression model, including all k explanatory variables, we impose that β^j is a subvector of β^f for all j 's. Therefore the models \mathfrak{M}_j and therefore the mixture model (8) all are parameterised in terms of *the same* β^f . To simplify the notation, we will denote this common parameter vector by $\beta = (\beta_0, \beta_1, \dots, \beta_k)^T$. Therefore, conditional on $\zeta_i = j$, we have

$$y_i \sim \mathcal{N}(X(i) \cdot j_2 \beta, \sigma^2),$$

where $X(i)$ denotes the i -th row of X and j_2 is the binary (base 2) representation of the integer j , with the convention that $X(i) \cdot j_2$ means a term-by-term multiplication, i.e.,

that this vector contains zero entries for the components of j_2 that are equal to zero:

$$X(i) \cdot j_2 = (X(i)_1 j_2[1], \dots, X(i)_k j_2[k]).$$

Assuming $v_j > 0$ and gathering all observations such that $\zeta_i = j$ under the notation $y_{i;\zeta_i=j}$ and the corresponding covariates by $X_{i;\zeta_i=j}$, we then have

$$y_{i;\zeta_i=j} \sim \mathcal{N}_{v_j} (X_{i;\zeta_i=j} \cdot j_2 \beta, \sigma^2 I_{v_j}),$$

with the same convention about the term-by-term multiplication. The overall model conditional on $\zeta = (\zeta_1, \dots, \zeta_n)$, the conditional distribution of the dataset, is therefore

$$\begin{pmatrix} y_{i;\zeta_i=1} \\ y_{i;\zeta_i=2} \\ \vdots \\ y_{i;\zeta_i=\gamma} \end{pmatrix}_{n \times 1} = \begin{pmatrix} \mathbf{1}_{i;\zeta_i=1} & 1_2[1][X_1]_{i;\zeta_i=1} & \dots & 1_2[k][X_k]_{i;\zeta_i=1} \\ \mathbf{1}_{i;\zeta_i=2} & 2_2[1][X_1]_{i;\zeta_i=2} & \dots & 2_2[k][X_k]_{i;\zeta_i=2} \\ \vdots & \vdots & & \vdots \\ \mathbf{1}_{i;\zeta_i=\gamma} & \gamma_2[1][X_1]_{i;\zeta_i=\gamma} & \dots & \gamma_2[k][X_k]_{i;\zeta_i=\gamma} \end{pmatrix}_{n \times (k+1)} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}_{(k+1) \times 1} + \epsilon_{n \times 1}$$

where $\mathbf{1}_{i;\zeta_i=j}$ is a v_j -dimensional vector of 1's. By convention, any value of j such that $v_j = 0$ does not appear in the above. If we summarise the above equation as $\mathbf{y}_\zeta = \mathbf{X}_\zeta \beta + \epsilon$ and use a Zellner's (1986) G -prior,

$$\beta | \sigma \sim \mathcal{N}_{k+1} (M_{k+1}, c\sigma^2 (X^T X)^{-1}), \quad \pi(\sigma^2) \propto 1/\sigma^2,$$

the full conditional posterior distribution on the parameters is defined as

$$(\alpha_1, \dots, \alpha_\gamma) | \zeta \sim \mathcal{D}_\gamma(v_1 + a_0, \dots, v_\gamma + a_0), \beta | y, \zeta, \sigma \sim \mathcal{N}_{k+1}(\bar{\beta}, \bar{\Sigma}), \sigma^2 | y, \beta \sim \mathcal{IG}(a, b),$$

where

$$\begin{aligned} \bar{\beta} &= \bar{\Sigma} \{X^T X M / c\sigma^2 + \mathbf{X}_\zeta^T \mathbf{y}_\zeta / \sigma^2\} \\ \bar{\Sigma} &= \{X^T X / c\sigma^2 + \mathbf{X}_\zeta^T \mathbf{X}_\zeta / \sigma^2\}^{-1} \\ a &= (n + k + 1)/2 \\ b &= (\mathbf{y}_\zeta - \mathbf{X}_\zeta \beta)^T (\mathbf{y}_\zeta - \mathbf{X}_\zeta \beta) / 2 + (\beta - M)^T (X^T X) (\beta - M) / 2c. \end{aligned}$$

The MCMC implementation of this version of the model then leads to a straightforward Gibbs sampler.

Case 2. The alternative parameterisation of the mixture (1) is to consider all regression coefficients as independent between models. This means that, for $j = 1, \dots, \gamma$, the regression model \mathfrak{M}_j is written as $y = X^j \beta_{\mathfrak{M}_j} + \epsilon$ and that the $\beta_{\mathfrak{M}_j}$'s are independent. We still assume σ is common to all components. In this representation, we allocate a Zellner's G -prior to each parameter vector,

$$\beta_{\mathfrak{M}_j} \sim \mathcal{N}_{v_j} (M_j, c\sigma^2 (\{X^j\}^T X^j)^{-1})$$

and, conditional on the allocation vector ζ , the full conditional posterior distributions are easily derived:

$$(\alpha_1, \dots, \alpha_\gamma) | \zeta \sim \mathcal{D}_\gamma(v_1 + a_0, \dots, v_\gamma + a_0), \beta_{\mathfrak{M}_j} | y, \sigma, \zeta \sim \mathcal{N}_{v_j}(\eta_j, \varphi_j), \sigma^2 | y, \beta \sim \mathcal{IG}(a, b),$$

where

$$\begin{aligned}\eta &= \varphi \left\{ \{X^j\}^T X^j M_j / c\sigma^2 + X_{i;\zeta_i=j}^j y_{i;\zeta_i=j} / \sigma^2 \right\} \\ \varphi &= \left\{ \{X^j\}^T X^j / c\sigma^2 + \{X_{i;\zeta_i=j}^j\}^T X_{i;\zeta_i=j}^j / \sigma^2 \right\}^{-1} \\ a &= (n + s) / 2 \\ b &= \frac{1}{2} \sum_{j=1}^{\gamma} c(y_{i;\zeta_i=j} - X_{i;\zeta_i=j}^j \beta_{\mathfrak{M}_j})^T (y_{i;\zeta_i=j} - X_{i;\zeta_i=j}^j \beta_{\mathfrak{M}_j}) \\ &\quad + c^{-1} (\beta_{\mathfrak{M}_j} - M_j)^T (\{X^j\}^T X^j) (\beta_{\mathfrak{M}_j} - M_j)\end{aligned}$$

where s is the total number of the regression coefficients of all models under comparison and where the indexing conventions are the same as in Case 1.

Comparison of the performance of the mixture approach in both cases is conducted via simulated data with $k = 3$ covariates, meaning that $(1 \leq i \leq n)$

$$\mathbb{E}[y_i \mid \beta, X] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

This setting thus involves 15 models to be compared (since the model in which the mean of the observations is zero is of no interest). The parameters used for the data simulation are $(\beta_0, \beta_1, \beta_2, \beta_3) = (2, -3, 0, 0)$, $\sigma = 1$, with X_1 , X_2 and X_3 simulated from $\mathcal{N}(0, 1)$, $\mathcal{B}(1, .5)$ and $\mathcal{U}(10, 11)$, respectively. We are seeking to identify the true regression model

$$\mathfrak{M}_2 : y_i = 2 - 3X_{i1} + \epsilon_i,$$

by running (Gibbs) mixture estimations algorithms.

Based on a single simulated dataset, Figure 15 summarises the results of those simulations by representing the convergence of the posterior medians of the true model weight in both cases as the sample size n increases. Comments that stem from these results are that

- ✓ all posterior medians of the true model weight α_2 converge to 1 when the sample size increases to $n = 10,000$, which means that the mixture procedure eventually supports \mathfrak{M}_2 against the other models;
- ✓ the impact of the prior modelling, i.e., of the value of a_0 is such that the convergence is faster when a_0 is smaller;
- ✓ even for small sample sizes, the posterior medians of α_2 are close to 1;
- ✓ the differences between both mixture parameterisations, i.e., Case 1 and Case 2, are negligible;
- ✓ for almost every sample size and prior hyperparameter, the method concludes that \mathfrak{M}_2 is likely to be more appropriate than the others.

The most interesting conclusion is therefore that using completely independent parameterisation between the components of the mixture does not induce a strong degradation in the performances of the method, although the convergence to 1 is slightly slower on the right hand side of Figure 15. Table 3 produces the posterior means of α_2 under different Dirichlet hyperparameters a_0 ; this shows a stronger difference only for $a_0 = 0.5$, which then appears as a less reliable upper bound.

In order to contrast with the classical Bayesian analysis of this model, we compare our posterior means of $1 - \alpha_2$ with the posterior probability of \mathfrak{M}_2 computed using a G-prior for the regression parameters in Figure 16. This graph shows that the convergence of $\log(1 - \mathbb{E}(\alpha_2 \mid y, X))$

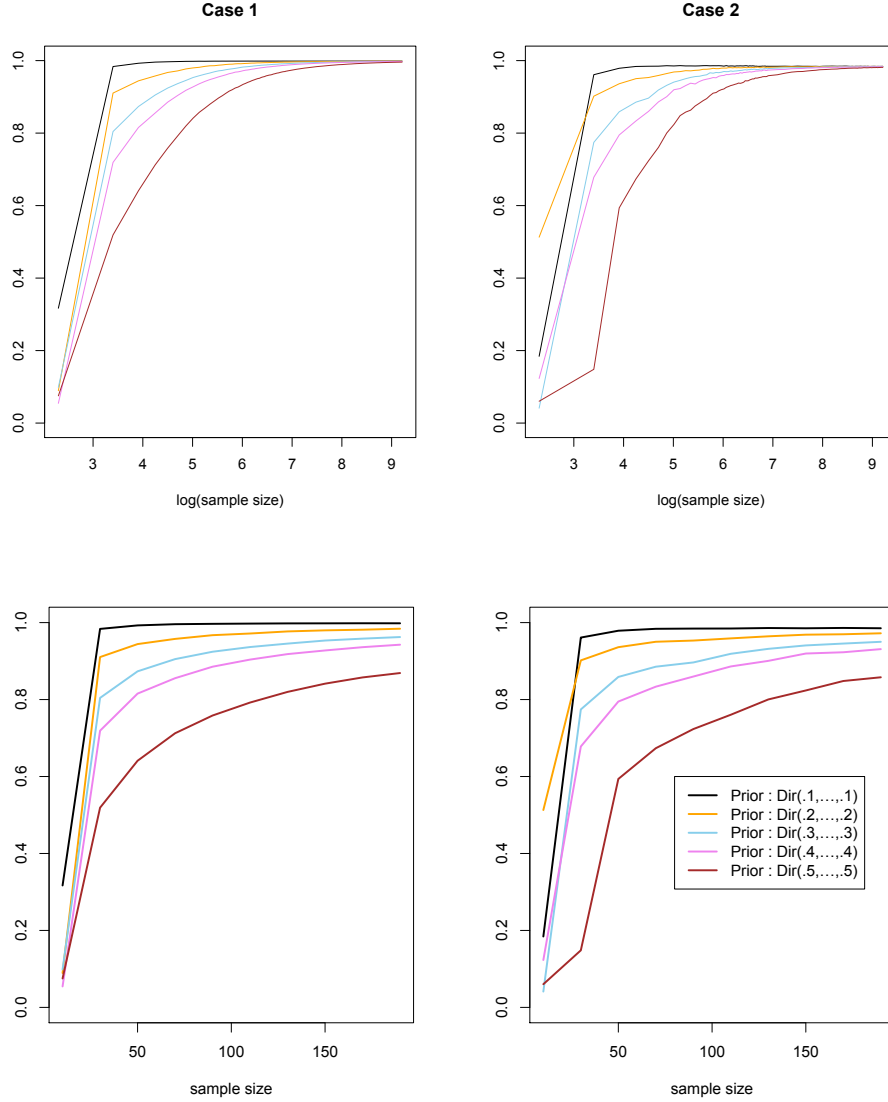


Fig 15: **Example 3.6:** (*top*) Posterior medians of the true model weight over five values of $e_0 = .1, .2, .3, .4, .5$ for sample sizes ranging from 1 to 10^4 and (*bottom*) from 1 to 200. Case 1 (*left*) and Case 2 (*right*) correspond to common and independent parameterisations of the mixture components. Each approximation is based on 10^4 Gibbs iterations.

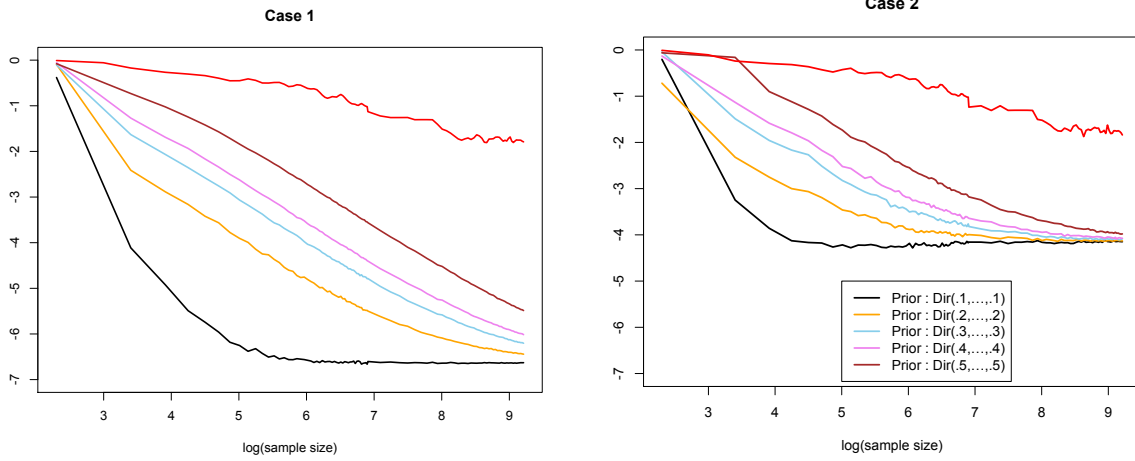
is faster than for $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$. It also highlights a difference between Cases 1 and 2 for the larger sample size, with $\log(1 - \mathbb{E}(\alpha_2|y, X))$ concentrated between -6.5 and -5 in Case 1 method, about -4 in Case 2, and about -2 for $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$. Although these figures are based on a single dataset, they are conclusive about the superior performance of the mixture approach.

As a second comparative evaluation of the mixture approach for linear models, for the same set of three regressors, we simulated 50 datasets with 500 observations from each of the models $\mathfrak{M}_1, \dots, \mathfrak{M}_{15}$ and examined the respective averages of the Bayes estimates and the posterior probabilities. In all cases reported in Table 4, much stronger support for the correct model is provided by the posterior means and medians than by the posterior probability.

TABLE 3

Example 3.6: Posterior means of α_2 , weight of model \mathfrak{M}_2 , when the sample size is $n=30$

| | | | | | |
|-----------------------------|--------|--------|--------|--------|--------|
| $a_0:$ | .1 | .2 | .3 | .4 | .5 |
| Case 1: | | | | | |
| $\mathbb{E}[\alpha_2 y, X]$ | 0.9836 | 0.9104 | 0.8043 | 0.7190 | 0.5190 |
| Case 2: | | | | | |
| $\mathbb{E}[\alpha_2 y, X]$ | 0.9611 | 0.9018 | 0.7743 | 0.6780 | 0.3905 |

Fig 16: **Example 3.6:** $\log(1 - \mathbb{E}(\alpha_2|y, X))$ and $\log(1 - \mathbb{P}(\mathfrak{M}_2|y, X))$ (red lines) over logarithm of the sample size for $a_0 = .1, .2, .3, .4, .5$. Each posterior approximation is based on 10^4 iterations.

4. CASE STUDY : A SURVIVAL ANALYSIS

Survival and reliability models are employed in a large number of disciplines ranging from engineering to health. An important modelling decision in these problems is the choice of the survival function. Among the many parametric alternatives, common choices include the Weibull, log-Normal, logistic, log-logistic, exponential, hypo- and hyper-exponential extensions, Gompertz, Birnbaum-Saunders, Erlang, Coxian, and Pareto distributions. The Weibull distribution is also a representative of the class of models used for extreme value modelling. Other models in this class include the extreme value, Stable, Gumbel and Fréchet distributions.

We apply here our testing paradigm to choosing between three potential survival models. Given data (x_1, \dots, x_n) with corresponding censoring indicators (c_1, \dots, c_n) , we wish to test the hypothesis that the data are drawn from a log-Normal(ϕ, κ^2), a Weibull(α, λ), or a log-Logistic(γ, δ) distribution. The corresponding mixture is thus given by the density

$$\alpha_1 \exp\{-(\log x - \phi)^2/2\kappa^2\}/\sqrt{2\pi}\kappa + \alpha_2 \frac{\alpha}{\lambda} \exp\{-(x/\lambda)^\alpha\}(x/\lambda)^{\alpha-1} + \alpha_3 (\delta/\gamma) (x/\gamma)^{\delta-1} / (1 + (x/\gamma)^\delta)^2$$

where $\alpha_1 + \alpha_2 + \alpha_3 = 1$. A more amenable version can be obtained by working on the scale $Y = -\log(X)$, which then provides a comparison between the $N(\theta, \sigma^2)$, Gumbel(μ, β), and Logistic(ξ, ζ) distributions. This gives rise to the mixture density

$$f_{\theta, \alpha}(y) = \alpha_1 \exp\{-(y - \phi)^2/2\sigma^2\}/(\sqrt{2\pi}\sigma) + \alpha_2^{1/\beta} \exp\{-(y - \mu)/\beta\} \exp\left\{-e^{-(y-\mu)/\beta}\right\} + \alpha_3 \exp\{-(y - \xi)/\zeta\} / \{\zeta(1 + \exp\{-(y - \xi)/\zeta\})^2\}.$$

TABLE 4

Example 3.6: Comparison between posterior probabilities of the true models, posterior means and medians of the mixture model weights, averaged over 50 replicas of samples of size 500.

| True model a_0 | $\mathbb{E}(\alpha_j y)$ | | | $\text{med}(\alpha_j y)$ | | | $\mathbb{P}(\mathfrak{M}_j y)$ |
|---------------------|--------------------------|------|------|--------------------------|------|------|--------------------------------|
| | .1 | .3 | .5 | .1 | .3 | .5 | |
| \mathfrak{M}_1 | .952 | .843 | .791 | 1 | 1 | .936 | .465 |
| \mathfrak{M}_2 | .983 | .962 | .786 | .989 | .994 | .915 | .411 |
| \mathfrak{M}_3 | .976 | .973 | .821 | 1 | 1 | .921 | .494 |
| \mathfrak{M}_4 | .991 | .867 | .902 | 1 | .987 | .934 | .503 |
| \mathfrak{M}_5 | .940 | .952 | .896 | .978 | .975 | .909 | .591 |
| \mathfrak{M}_6 | .974 | .939 | .898 | 1 | 1 | .940 | .617 |
| \mathfrak{M}_7 | .973 | .899 | .906 | 1 | 1 | 1 | .888 |
| \mathfrak{M}_8 | .991 | .918 | .924 | 1 | 1 | 1 | .938 |
| \mathfrak{M}_9 | .953 | .940 | .878 | 1 | .993 | .956 | .505 |
| \mathfrak{M}_{10} | .951 | .967 | .849 | .988 | .988 | .947 | .663 |
| \mathfrak{M}_{11} | .958 | .951 | .820 | 1 | .989 | .971 | .099 |
| \mathfrak{M}_{12} | .969 | .964 | .951 | .995 | .967 | .943 | .196 |
| \mathfrak{M}_{13} | .919 | .951 | .872 | 1 | .962 | .926 | .547 |
| \mathfrak{M}_{14} | .952 | .964 | .890 | .998 | .981 | .911 | .126 |
| \mathfrak{M}_{15} | .991 | .991 | .955 | 1 | .994 | .908 | .164 |

If we opt for a *common* parameterisation of those different models, we have the following moment matching equations

$$\begin{aligned}\phi &= \mu + \gamma\beta = \xi \\ \sigma^2 &= \pi^2\beta^2/6 = \zeta^2\pi^2/3\end{aligned}$$

where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant. As above, this choice allows the use of a non-informative prior on the common location scale parameter, $\pi(\phi, \sigma^2) = 1/\sigma^2$. Once again, we use a Dirichlet prior $\mathcal{D}(a_0, a_0, a_0)$ on $(\alpha_1, \alpha_2, \alpha_3)$. Appendix 6 establishes that the corresponding posterior is proper provided the observations are not all equal.

A common feature in survival data is the presence of censoring. In this case, the mixture equation becomes

$$\begin{aligned}f_{\theta,\alpha}(y, c) &= \alpha_1 \left[e^{-(y-\phi)^2/2\sigma^2} / \sqrt{2\pi}\sigma \right]^c \Phi[(y-\phi)/\sigma]^{1-c} + \\ &\alpha_2 \left[(1/\beta) e^{-(y-\mu)/\beta} \exp \left\{ -e^{-(y-\mu)/\beta} \right\} \right]^c \left[\exp \left\{ -e^{-(y-\mu)/\beta} \right\} \right]^{1-c} + \\ &\alpha_3 \left[e^{-(y-\xi)/\zeta} / \left\{ \zeta(1 + e^{-(y-\xi)/\zeta})^2 \right\} \right]^c \left[1 / \left\{ (1 + e^{-(y-\xi)/\zeta}) \right\} \right]^{1-c}.\end{aligned}$$

Two simulations experiments were performed to evaluate the mixture estimation approach in this context. First, the performance of the approach in distinguishing between the Weibull, lognormal and log-logistic distributions was assessed by simulating 1000 observations from a Normal(0, 1) density (with no censoring), and testing a Normal versus Gumbel and Logistic distributions as described above. The experiment was then repeated using 1000 simulations from a Gumbel and then from a Logistic distribution. For illustration, the moment-matched Normal, Gumbel and Logistic densities are depicted in the left hand panel of Figure 17.

The Gibbs sampler was run for 10,000 iterations using a prior value of $a_0 = 1.0$ for the hyperparameter on the mixture weights. The resultant probabilities of selecting the various distributions are shown in Figure 17, right panel. It can be seen that in all cases, the correct model was overwhelmingly identified. As expected, the probabilities of a correct selection increase with the sample size; in an analogous experiment with $n = 10^5$, all probabilities were larger than 0.90 (figures not shown).

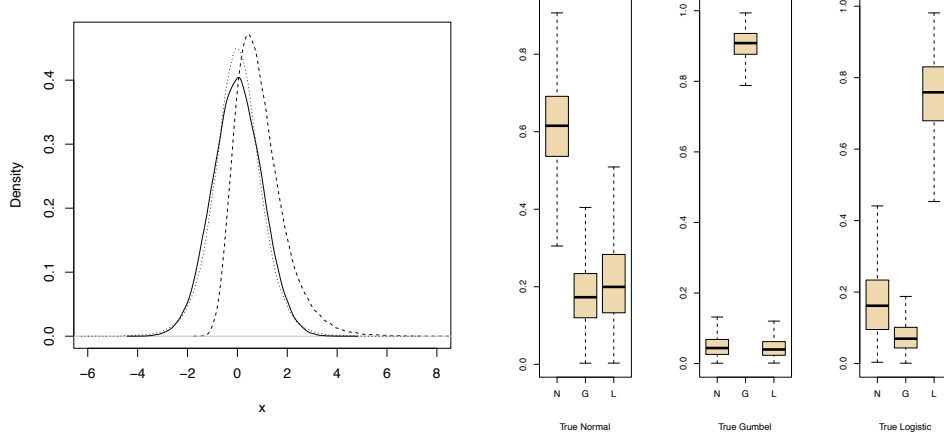


Fig 17: **Case study:** (left) Normal (solid), Gumbel (dashed) and Logistic (dotted) densities with $(0, 1)$ parameter; (right) Boxplots of the posterior distributions of the weights for the Normal, Gumbel and Logistic densities, respectively, for each of three scenarios: truth = Normal (left panel), truth = Gumbel (middle panel), truth = logistic (right panel).

The second simulation study was designed to assess the influence of the hyperparameter, a_0 . The above Monte Carlo experiment was repeated with $n = 1000$ simulated observations, comparing the impact of four values of a_0 , namely $a_0 = 0.01, 0.1, 1.0, 10.0$, for all three distributions and for each pair of distributions. Figure 18 and 19 depict the probabilities of selecting a (true) Normal or (true) Gumbel model. In agreement with earlier comparisons, the value of a_0 impacts on the probability of a correct model selection; however, in all cases covered by this Figure, the correct model was overwhelmingly identified as the most likely one. Note further that in this experiment the values of a_0 were larger than those we recommended above, namely $a + 0 \leq 0.5$. As before, increasing the sample size from $n = 1,000$ to $n = 10,000$ pushes the posterior probabilities toward the boudaries.

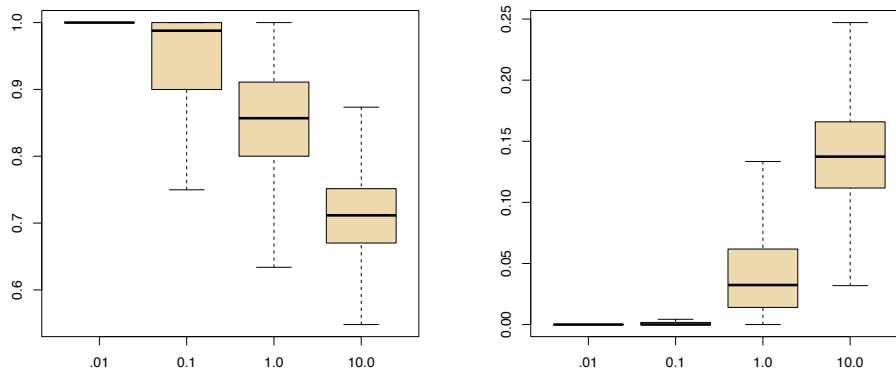


Fig 18: **Case study:** Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii: truth = Normal (left panel), truth = Gumbel (right panel), $a_0 = 0.01, 0.1, 1.0, 10.0$ (from left to right in each panel) and $n = 1,000$ observations.

In addition to these Monte Carlo evaluations of the mixture approach, we considered a real

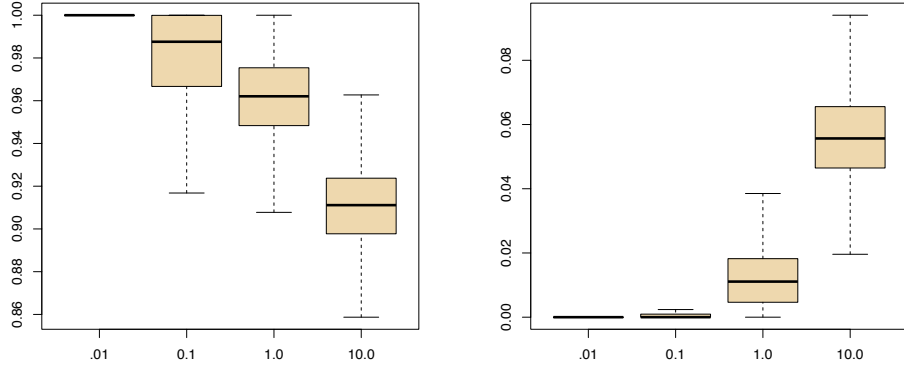


Fig 19: **Case study:** Boxplots of the posterior distributions of the Normal weight α_1 under the two scenarii: truth = Normal (*left panel*), truth = Gumbel (*right panel*), $a_0=0.01, 0.1, 1.0, 10.0$ (from left to right in each panel) and $n = 10,000$ simulated observations.

case study involving modelling survival times for breast cancer in Queensland, Australia. A sample of 25,125 individuals with breast cancer was provided by Cancer Council Queensland. Among the subjects, 83.5% were recorded as censored and the remainder ($n = 4155$) were recorded as deaths from any cause. The median survival times were 4.35 and 2.02 years for each of these groups, respectively.

The response variable used in the following analyses is the hazard function, defined as the probability of death at $t + \delta$ years given survival to t years, adjusted for age, sex and the expected mortality rate, that is, the age- and sex-adjusted background population risk of death. Of interest is whether or not this distribution is best fitted by a log Normal, Weibull, or log-Logistic distribution, or, equivalently, whether or not the log hazard is best fitted by a Normal, Gumbel, or Logistic distribution. These three sets of distributions have different fit characteristics: whereas the Normal (and hence the log Normal) distribution fits the centre of the distribution more closely, the Weibull (and hence the Gumbel) distribution captures the tail behaviour more accurately. The logistic distribution had a similar fit to the Normal, but it accommodates slightly more diffuse tails.

As an alternative to selecting a single model, it may be preferable to follow a model averaged approach. This is intrinsically part of the mixture model approach since the MCMC outcome provides in addition a posterior approximation of the overall mixture. (It could actually be argued that this approach is even better than standard model averaging as each observation in the sample selects the best fitted component of the mixture.) The choice of an appropriate model or of a combination of models is important for the prediction of survival for cancer patients, which then impacts on decisions about personalised management and treatment options.

Based on a choice of hyperparameter $a_0 = 1.0$, the mixture test for the breast cancer data resulted in the choice of the logistic distribution with probability 0.996 (s.d. $1.4 \cdot 10^{-3}$), with the remaining probability mass almost equally split between the Normal and Gumbel distributions.

5. ASYMPTOTIC CONSISTENCY

In this section we prove posterior consistency for our mixture testing procedure. More precisely we study the asymptotic behaviour of the posterior distribution of α . We consider two different cases. In the first case, the two models, \mathfrak{M}_1 and \mathfrak{M}_2 , are well separated while, in

the second case, model \mathfrak{M}_1 is a submodel of \mathfrak{M}_2 . We denote by π the prior distribution on $(\alpha, \theta_1, \theta_2)$ and assume that $\theta_j \in \Theta_j \subset \mathbb{R}^{d_j}$. We first prove that, under weak regularity conditions on each model, we can obtain posterior concentration rates for the marginal density $f_{\theta, \alpha}(\cdot) = \alpha f_{1, \theta_1}(\cdot) + (1 - \alpha) f_{2, \theta_2}(\cdot)$. Let $\mathbf{x}^n = (x_1, \dots, x_n)$ be a n sample with true density f^* .

Proposition 1 *Assume that, for all $C_1 > 0$, there exist Θ_n a subset of $\Theta_1 \times \Theta_2$ and $B > 0$ such that*

$$(9) \quad \pi[\Theta_n^c] \leq n^{-C_1}, \quad \Theta_n \subset \{\|\theta_1\| + \|\theta_2\| \leq n^B\}$$

and that there exist $H \geq 0$ and $L, \delta > 0$ such that, for $j = 1, 2$,

$$(10) \quad \sup_{\theta, \theta' \in \Theta_n} \|f_{j, \theta_j} - f_{j, \theta'_j}\|_1 \leq L n^H \|\theta_j - \theta'_j\|, \quad \theta = (\theta_1, \theta_2), \theta' = (\theta'_1, \theta'_2),$$

$$\forall \|\theta_j - \theta_j^*\| \leq \delta; \quad KL(f_{j, \theta_j}, f_{j, \theta_j^*}) \lesssim \|\theta_j - \theta_j^*\|.$$

We then have that, when $f^ = f_{\theta^*, \alpha^*}$, with $\alpha^* \in [0, 1]$, there exists $M > 0$ such that*

$$\pi \left[(\alpha, \theta); \|f_{\theta, \alpha} - f^*\|_1 > M \sqrt{\log n / n} | \mathbf{x}^n \right] = o_p(1).$$

The proof of Proposition 1 is a direct consequence of Theorem 2.1 of Ghosal et al. (2000) and is omitted for the sake of conciseness. Condition (10) is a weak regularity condition on each of the candidate models. Combined with condition (9) it allows to consider non-compact parameter sets in the usual way, see for instance Ghosal et al. (2000). It is satisfied in all examples considered in Section 3. We build on Proposition 1 to describe the asymptotic behaviour of the posterior distribution on the parameters.

5.1 The case of separated models

Assume that both models are separated in the sense that there is identifiability:

$$(11) \quad \forall \alpha, \alpha' \in [0, 1], \quad \forall \theta_j, \theta'_j, j = 1, 2 \quad P_{\theta, \alpha} = P_{\theta', \alpha'} \Rightarrow \alpha = \alpha', \quad \theta = \theta',$$

where $P_{\theta, \alpha}$ denotes the distribution associated with $f_{\theta, \alpha}$. We assume that (11) also holds on the boundary of $\Theta_1 \times \Theta_2$. In other words, the following

$$\inf_{\theta_1 \in \Theta_1} \inf_{\theta_2 \in \Theta_2} \|f_{1, \theta_1} - f_{2, \theta_2}\|_1 > 0$$

holds. We also assume that, for all $\theta_j^* \in \Theta_j$, $j = 1, 2$, if P_{θ_j} converges in the weak topology to $P_{\theta_j^*}$, then θ_j converges in the Euclidean topology to θ_j^* . The following result then holds:

Theorem 1 *Assume that (11) is satisfied, together with (9) and (10), then for all $\epsilon > 0$*

$$\pi[|\alpha - \alpha^*| > \epsilon | \mathbf{x}^n] = o_p(1).$$

In addition, assume that the mapping $\theta_j \rightarrow f_{j, \theta_j}$ is twice continuously differentiable in a neighbourhood of θ_j^ , $j = 1, 2$, and that*

$$f_{1, \theta_1^*} - f_{2, \theta_2^*}, \nabla f_{1, \theta_1^*}, \nabla f_{2, \theta_2^*}$$

are linearly independent as functions of y and that there exists $\delta > 0$ such that

$$\nabla f_{1, \theta_1^*}, \nabla f_{2, \theta_2^*}, \sup_{|\theta_1 - \theta_1^*| < \delta} |D^2 f_{1, \theta_1}|, \sup_{|\theta_2 - \theta_2^*| < \delta} |D^2 f_{2, \theta_2}| \in L_1.$$

Then

$$(12) \quad \pi[|\alpha - \alpha^*| > M \sqrt{\log n / n} | \mathbf{x}^n] = o_p(1).$$

Theorem 1 allows for the interpretation of the quantity α under the posterior distribution. In particular, if the data \mathbf{x}^n is generated from model \mathfrak{M}_1 (resp. \mathfrak{M}_2), then the posterior distribution on α concentrates around $\alpha = 1$ (resp. around $\alpha = 0$), which establishes the consistency of our mixture approach.

Proof: Using Proposition 1, we have that

$$\pi(A_n|\mathbf{x}^n) = 1 + o_p(1)$$

with $A_n = \{(\alpha, \theta); \|f_{\theta, \alpha} - f_{\theta^*, \alpha^*}\|_1 \leq \delta_n\}$ and $\delta_n = M\sqrt{\log n/n}$. Consider a subsequence $\alpha_n, P_{1, \theta_{1n}}, P_{2, \theta_{2n}}$ which converges to α, μ_1, μ_2 where convergence holds in the sense that $\alpha_n \rightarrow \alpha$ and $P_{j, \theta_{jn}}$ converges weakly to μ_j . Note that $\mu_j(\mathcal{X}) \leq 1$ by precompactity of the unit ball under the weak topology. At the limit

$$\alpha\mu_1 + (1 - \alpha)\mu_2 = \alpha^*P_{1, \theta_1^*} + (1 - \alpha^*)P_{2, \theta_2^*}$$

The above equality implies that μ_1 and μ_2 are probabilities. Using (11), we obtain that

$$\alpha = \alpha^*, \quad \mu_j = P_{j, \theta_j^*},$$

which implies posterior consistency for α . The proof of (12) follows the same line as in [Rousseau and Mengersen \(2011\)](#). Consider first the case where $\alpha^* \in (0, 1)$. Then the posterior distribution on θ concentrates around θ^* .

Writing

$$L' = (f_{1, \theta_1^*} - f_{2, \theta_2^*}, \alpha^* \nabla f_{1, \theta_1^*}, (1 - \alpha^*) \nabla f_{2, \theta_2^*}) := (L_\alpha, L_1, L_2)$$

$$L'' = \text{diag}(0, \alpha^* D^2 f_{1, \theta_1^*}, (1 - \alpha^*) D^2 f_{2, \theta_2^*}) \quad \text{and} \quad \eta = (\alpha - \alpha^*, \theta_1 - \theta_1^*, \theta_2 - \theta_2^*), \quad \omega = \eta/|\eta|,$$

we then have

$$(13) \quad \|f_{\theta, \alpha} - f_{\theta^*, \alpha^*}\|_1 = |\eta| \left| \omega^T L' + |\eta|/2 \omega^T L'' \omega + |\eta| \omega_1 \left[\omega_2 L'_2 + \omega_3 L'_3 \right] + o(|\eta|) \right|$$

For all $(\alpha, \theta) \in A_n$, set $\eta = (\alpha - \alpha^*, \theta_1 - \theta_1^*, \theta_2 - \theta_2^*)$ goes to 0 and for n large enough there exists $\epsilon > 0$ such that $|\alpha - \alpha^*| + |\theta - \theta^*| \leq \epsilon$. We now prove that there exists $c > 0$ such that for all $(\alpha, \theta) \in A_n$

$$v(\omega) = \left| \omega^T L' + \frac{|\eta|}{2} \omega^T L'' \omega + |\eta| \omega_1 \left[\omega_2^T L'_2 + \omega_3^T L'_3 \right] + o(|\eta|) \right| > c,$$

where ω is defined with respect to α, θ . Were it not the case, there would exist a sequence $(\alpha_n, \theta_n) \in A_n$ such that the associated $v(\omega_n) \leq c_n$ with $c_n = o(1)$. As ω_n belongs to a compact set we could find a subsequence converging to a point $\bar{\omega}$. At the limit we would obtain

$$\bar{\omega}^T L' = 0$$

and by linear independence $\bar{\omega} = 0$ which is not possible. Thus for all $(\alpha, \theta) \in A_n$

$$|\alpha - \alpha^*| + |\theta - \theta^*| \lesssim \delta_n.$$

Assume now instead that $\alpha^* = 0$. Then define $L' = (L_\alpha, L_2)$ and

$$L'' = \text{diag}(0, D^2 f_{2, \theta_2^*}) \quad \text{and} \quad \eta = (\alpha - \alpha^*, \theta_2 - \theta_2^*), \quad \omega = \eta/|\eta|$$

and consider a Taylor expansion with θ_1 fixed, $\theta_1^* = \theta_1$ and $|\eta|$ going to 0. This leads to

$$(14) \quad \|f_{\theta, \alpha} - f_{\alpha^*, \theta^*}\|_1 = |\eta| \left| \omega^T L' + \frac{|\eta|}{2} \omega^T L'' \omega + |\eta| \omega_1 \omega_3 L'_3 \right| + o(|\eta|)$$

in place of (13) and the posterior concentration rate δ_n is obtained in the same way. \square

We now consider the embedded case.

5.2 Embedded case

In this section we assume that \mathfrak{M}_1 is a submodel of \mathfrak{M}_2 , in the sense that $\theta_2 = (\theta_1, \psi)$ with $\psi \in \mathcal{S} \subset \mathbb{R}^d$ and that $f_{2,\theta_2} \in \mathfrak{M}_1$ when $\theta_2 = (\theta_1, \psi_0)$ for some given value ψ_0 , say $\psi_0 = 0$. Condition (11) is no longer verified for all α 's: we assume however that it is verified for all $\alpha, \alpha^* \in (0, 1]$ and that $\theta_2^* = (\theta_1^*, \psi^*)$ satisfies $\psi^* \neq 0$. In this case, under the same conditions as in Theorem 1, we immediately obtain the posterior concentration rate of $\sqrt{\log n/n}$ for estimating α when $\alpha^* \in (0, 1)$ and $\psi^* \neq 0$. We now treat the case where $\psi^* = 0$; in other words, f^* is in model \mathfrak{M}_1 .

As in Rousseau and Mengersen (2011), we consider both possible paths to approximate f^* : either α goes to 1 or ψ goes to $\psi_0 = 0$. In the first case, called path 1, $(\alpha^*, \theta^*) = (1, \theta_1^*, \theta_1^*, \psi)$ with $\psi \in \mathcal{S}$. In the second case, called path 2, $(\alpha^*, \theta^*) = (\alpha, \theta_1^*, \theta_1^*, 0)$ with $\alpha \in [0, 1]$. In both cases, denote the distributions by P^* . We also denote $F^*g = \int f^*(x)g(x)d\mu(x)$ for any integrable function g . For sparsity reasons, we consider the following structure for the prior on (α, θ) :

$$\pi(\alpha, \theta) = \pi_\alpha(\alpha)\pi_1(\theta_1)\pi_\psi(\psi), \quad \theta_2 = (\theta_1, \psi).$$

This means that the parameter θ_1 is common to both models, i.e., that θ_2 shares the parameter θ_1 with f_{1,θ_1} .

Condition (11) is replaced by

$$(15) \quad P_{\theta,\alpha} = P^* \Rightarrow \alpha = 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, \psi) \quad \text{or} \quad \alpha \leq 1, \quad \theta_1 = \theta_1^*, \quad \theta_2 = (\theta_1^*, 0)$$

Let Θ^* be the parameter set corresponding to P^* .

As in the case of separated models, the posterior distribution concentrates on Θ^* . We now describe more precisely the asymptotic behaviour of the posterior distribution, using Theorem 1 of Rousseau and Mengersen (2011). Since this theorem cannot be applied directly, we adapt it as follows. We require the following assumptions with $f^* = f_{1,\theta_1^*}$. For the sake of simplicity, we assume that Θ_1 and \mathcal{S} are compact. Extensions to non-compact sets can be handled similarly to Rousseau and Mengersen (2011).

B1 *Regularity*: Assume that $\theta_1 \rightarrow f_{1,\theta_1}$ and $\theta_2 \rightarrow f_{2,\theta_2}$ are 3 times continuously differentiable and that

$$\begin{aligned} F^* \left(\frac{\bar{f}_{1,\theta_1^*}^3}{\underline{f}_{1,\theta_1^*}^3} \right) &< +\infty, \quad \bar{f}_{1,\theta_1^*} = \sup_{|\theta_1 - \theta_1^*| < \delta} f_{1,\theta_1}, \quad \underline{f}_{1,\theta_1^*} = \inf_{|\theta_1 - \theta_1^*| < \delta} f_{1,\theta_1} \\ F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} |\nabla f_{1,\theta_1^*}|^3}{\underline{f}_{1,\theta_1^*}^3} \right) &< +\infty, \quad F^* \left(\frac{|\nabla f_{1,\theta_1^*}|^4}{\underline{f}_{1,\theta_1^*}^4} \right) < +\infty, \\ F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} |D^2 f_{1,\theta_1^*}|^2}{\underline{f}_{1,\theta_1^*}^2} \right) &< +\infty, \quad F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} |D^3 f_{1,\theta_1^*}|}{\underline{f}_{1,\theta_1^*}} \right) < +\infty \end{aligned}$$

B2 *Integrability*: There exists $\mathcal{S}_0 \subset \mathcal{S} \cap \{|\psi| > \delta_0\}$, for some positive δ_0 and satisfying $\text{Leb}(\mathcal{S}_0) > 0$, such that for all $\psi \in \mathcal{S}_0$,

$$F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f_{2,\theta_1,\psi}}{\underline{f}_{1,\theta_1^*}^4} \right) < +\infty, \quad F^* \left(\frac{\sup_{|\theta_1 - \theta_1^*| < \delta} f_{2,\theta_1,\psi}^3}{\underline{f}_{1,\theta_1^*}^3} \right) < +\infty,$$

B3 *Stronger identifiability* : Set

$$\nabla f_{2,\theta_1^*,\psi^*}(x) = (\nabla_{\theta_1} f_{2,\theta_1^*,\psi^*}(x)^\top, \nabla_{\psi} f_{2,\theta_1^*,\psi^*}(x)^\top)^\top.$$

Then for all $\psi \in \mathcal{S}$ with $\psi \neq 0$, if $\eta_0 \in \mathbb{R}$, $\eta_1 \in \mathbb{R}^{d_1}$

$$(16) \quad \eta_0(f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}) + \eta_1^\top \nabla_{\theta_1} f_{1,\theta_1^*} = 0 \quad \Leftrightarrow \eta_1 = 0, \eta_2 = 0$$

We can now state the main theorem:

Theorem 2 *Given the model*

$$f_{\theta_1,\psi,\alpha} = \alpha f_{1,\theta_1} + (1 - \alpha) f_{2,\theta_1,\psi},$$

assume that the data comprises the n sample $\mathbf{x}^n = (x_1, \dots, x_n)$ generated from f_{1,θ_1^*} for some $\theta_1^* \in \Theta_1$, that assumptions B1 – B3 are satisfied, and that there exists $M > 0$ such that

$$\pi \left[(\alpha, \theta); \|f_{\theta,\alpha} - f^*\|_1 > M \sqrt{\log n/n} |\mathbf{x}^n| \right] = o_p(1).$$

If the prior π_α on α is a Beta $\mathcal{B}(a_1, a_2)$ distribution, with $a_2 < d_2$, and if the prior $\pi_{\theta_1,\psi}$ is absolutely continuous with positive and continuous density at $(\theta_1^*, 0)$, then for all M_n going to infinity,

$$(17) \quad \pi \left[|\alpha - \alpha^*| > M_n (\log n)^\gamma / \sqrt{n} |\mathbf{x}^n| \right] = o_p(1), \quad \gamma = \max((d_1 + a_2)/(d_2 - a_2), 1)/2,$$

Proof: We must find a precise lower bound on

$$D_n := \int_{\alpha} \int_{\Theta} e^{l_n(f_{\theta,\alpha}) - l_n(f^*)} d\pi_{\theta}(\theta) d\pi_{\alpha}(\alpha)$$

Consider the approximating set

$$S_n(\epsilon) = \{(\theta, \alpha), \alpha > 1 - 1/\sqrt{n}, |\theta_1 - \theta_1^*| \leq 1/\sqrt{n}, |\psi - \bar{\psi}| \leq \epsilon\}$$

with $|\bar{\psi}| > 2\epsilon$ some fixed parameter in \mathcal{S} . Using the same computations as in [Rousseau and Mengersen \(2011\)](#), it holds that for all $\delta > 0$ there exists $C_\delta > 0$ such that

$$(18) \quad P^* \left(D_n < e^{-C_\delta} \pi(S_n(\epsilon))/2 \right) < \delta.$$

Hence with probability greater than $1 - \delta$, $D_n \gtrsim n^{-(b+d_1)/2}$. Denote $B_n = \{(\theta, \alpha); \|f_{\theta,\alpha} - f^*\|_1 \leq M \sqrt{\log n/n}\}$ and $A_n = \{(\theta, \alpha) \in B_n; 1 - \alpha > z_n/\sqrt{n}\}$ with $z_n = M_n (\log n)^\gamma / \sqrt{n}$ and M_n a sequence increasing to infinity. We split B_n into

$$B_{n,1}(\epsilon) = B_n \cap \{(\theta, \alpha), \theta = (\theta_1, \psi); |\psi| < \epsilon\}, \quad B_{n,2}(\epsilon) = B_n \cap B_{n,1}(\epsilon)^c.$$

To prove Theorem 2 it is enough to verify that

$$\pi(A_n) = o(n^{-(a_2+d_1)/2}).$$

To simplify notation we also write $\delta_n = M \sqrt{\log n/n}$. First we prove that for all $\epsilon > 0$, $A_n \cap B_{n,2}(\epsilon) = \emptyset$, when n is large enough. Let $\epsilon > 0$, then for any $(\theta, \alpha) \in A_n \cap B_{n,2}(\epsilon)$, We thus

have $|\psi| \neq o(1)$, $\alpha = 1 + o(1)$ and $|\theta_1 - \theta_1^*| = o(1)$. Consider a Taylor expansion of $f_{\theta,\alpha}$ around $\alpha = 1$ and $\theta_1 = \theta_1^*$, with ψ fixed. This leads to

$$\begin{aligned} f_{\theta,\alpha} - f^* &= (\alpha - 1)[f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}] + (\theta_1 - \theta_1^*)\nabla_{\theta_1} f_{1,\theta_1^*} + \frac{1}{2}(\theta_1 - \theta_1^*)^T (\bar{\alpha} f_{1,\bar{\theta}_1} + (1 - \bar{\alpha}) f_{2,\bar{\theta}_1,\psi}) (\theta_1 - \theta_1^*) \\ &\quad + (\alpha - 1)(\theta_1 - \theta_1^*)^T [\nabla_{\theta_1} f_{1,\bar{\theta}_1} - \nabla_{\theta_1} f_{2,\bar{\theta}_1,\psi}] \\ &= (\alpha - 1)[f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}] + (\theta_1 - \theta_1^*)\nabla_{\theta_1} f_{1,\theta_1^*} + o(|\alpha - 1| + |\theta_1 - \theta_1^*|) \end{aligned}$$

with $\bar{\alpha} \in (0, 1)$ and $\bar{\theta}_1 \in (\theta_1, \theta_1^*)$ and the $o(1)$ is uniform over $A_n \cap B_{n,2}(\epsilon)$. Set $\eta = (\alpha - 1, \theta_1 - \theta_1^*)$ and $x = \eta/|\eta|$ if $|\eta| > 0$. Then

$$\|f_{\theta,\alpha} - f^*\|_1 = |\eta| (x^T L_1(\psi) + o(1)), \quad L_1 = (f_{1,\theta_1^*} - f_{2,\theta_1^*,\psi}, \nabla_{\theta_1} f_{1,\theta_1^*})$$

We now prove that on $A_n \cap B_{n,2}(\epsilon)$, $\|f_{\theta,\alpha} - f^*\|_1 \gtrsim |\eta|$. Assume that it is not the case: then there exist $c_n > 0$ going to 0 and a sequence (θ_n, α_n) such that along that subsequence $|x_n^T L_1(\psi_n) + o(1)| \leq c_n$ with $x_n = \eta_n/|\eta_n|$. Since it belongs to a compact set, together with ψ_n , any converging subsequence satisfies at the limit $(\bar{x}, \bar{\psi})$,

$$\bar{x}^T L_1(\bar{\psi}) = 0,$$

which is not possible. Hence $|\alpha - 1| \lesssim M\sqrt{\log n}/\sqrt{n} = o(M_n(\log n)^\gamma/\sqrt{n})$, which is not possible so that $A_n \cap B_{n,2}(\epsilon) = \emptyset$ when n is large enough. We now bound $\pi(A_n \cap B_{n,1}(\epsilon))$ for $\epsilon > 0$ small enough but fixed. We consider a Taylor expansion around $\theta^* = (\theta_1^*, 0)$, leaving α fixed. Note that $\nabla_{\theta_1} f_{2,\theta^*} = \nabla_{\theta_1} f_{1,\theta_1^*}$. We have

$$f_{\theta,\alpha} - f^* = (\theta_1 - \theta_1^*)^T \nabla_{\theta_1} f_{2,\theta^*} + (1 - \alpha)\psi^T \nabla_\psi f_{2,\theta^*} \frac{1}{2}(\theta - \theta^*)^T H_{\alpha,\bar{\theta}}(\theta - \theta^*)$$

where $H_{\alpha,\bar{\theta}}$ is the block matrix

$$H_{\alpha,\bar{\theta}} = \begin{pmatrix} \alpha D_{\theta_1}^2 f_{1,\bar{\theta}_1} + (1 - \alpha) D_{\theta_1,\theta_1}^2 f_{2,\bar{\theta}} & (1 - \alpha) D_{\theta_1,\psi}^2 f_{2,\bar{\theta}} \\ (1 - \alpha) D_{\psi,\theta_1}^2 f_{2,\bar{\theta}} & (1 - \alpha) D_{\psi,\psi}^2 f_{2,\bar{\theta}} \end{pmatrix}.$$

Since $H_{\alpha,\bar{\theta}}$ is bounded in L_1 (in the sense that each of its components is bounded as functions in L_1), uniformly in neighbourhoods of θ^* , then writing $\eta = (\theta_1 - \theta_1^*, (1 - \alpha)\psi)$ and $x = \eta/|\eta|$, we have that $|\eta| = o(1)$ on $A_n \cap B_{n,1}(\epsilon)$ and

$$\|f_{\theta,\alpha} - f^*\|_1 \gtrsim |\eta| (x^T \nabla f_{2,\theta^*} + o(1)),$$

if ϵ is small enough. Using a similar argument to before, this leads to $|\eta| \lesssim \delta_n$ on $A_n \cap B_{n,1}(\epsilon)$, so that

$$\pi(A_n \cap B_{n,1}(\epsilon)) \lesssim \delta_n^{d_1} \int_{z_n/\sqrt{n}}^1 (\delta_n/u)^{d_2} u^{b-1} du \lesssim \delta_n^{d_1+b} z_n^{b-d_2} \lesssim n^{-(d_1+a_2)/2} M_n^{a_2-d_2},$$

which concludes the proof. \square

6. CONCLUSION

Bayesian inference has been used in a very wide range over the past twenty years, mostly thanks to enhanced computing abilities, and many of those applications of the Bayesian paradigm have concentrated on the comparison of scientific theories and on testing of hypotheses. Due to the ever increasing complexity of the statistical models handled in such applications, the

natural and understandable tendency of practitioners has been to rely on the posterior probability or the Bayes factor. Very often, warnings about the validity of these approaches have gone unheeded (Robert et al., 2011), in particular concerning the poorly understood sensitivity of such tools to both prior modelling and posterior calibration. In this space, objective Bayes solutions remain tentative and do not meet with consensus.

We thus believe Bayesian analysis has reached the time for a paradigm shift in the matter of hypothesis testing and model selection. Importantly, the solution does not have to be found outside the Bayesian paradigm, as for instance the frequentist priors of Johnson (2013b,a) and the integrated likelihood setting of Aitkin (2010). The novel paradigm we proposed here for Bayesian testing of hypotheses and Bayesian model comparison offers many incentives while answering some of the classical attacks against posterior probabilities and Bayes factors. Our alternative to the construction of traditional posterior probabilities of a given hypothesis or the assertion that the data originate from a specific model is therefore to rely on the encompassing mixture model. Not only do we replace the original testing problem with a better controlled estimation target that focuses on the probability of a given model within the mixture model, but we also allow for posterior variability over this frequency. This is in contrast to the deterministic characteristics of the standard Bayesian approach. The posterior distribution on the weights of both components in the mixture offers a setting for deciding about which model is most favoured by the data that is at least as intuitive as the sole number corresponding to either the posterior probability or the Bayes factor. The range of acceptance, rejection and indecision conclusions can easily be calibrated by simulation under both models, as well as by deciding on the values of the weights that are extreme enough in favour of one model. The examples provided in this paper have shown that if one model is indeed correct, the posterior medians of the corresponding weight in the mixture settles very quickly near the boundary values of 1. Although we do not advocate such practice, it is even possible to derive a Bayesian p -value by considering the posterior area under the tail of the distribution of the weight.

Besides decision making, another issue of potential concern about this new approach is the impact of the prior modelling. We demonstrated through all our examples that a partly common parameterisation is always feasible and hence allows for reference priors, at least on the common parameters. This proposal thus allows for removal of the absolute prohibition of using improper priors in hypothesis testing (DeGroot, 1973), a problem which has plagued the objective Bayes literature for decades. Concerning the prior on the weight parameter, we analysed the sensitivity on the resulting posterior distribution of various prior Beta distributions on those weights. While the sensitivity is clearly present, it naturally vanishes as the sample size increases, in agreement with our consistency results, and remains of a moderate magnitude. This leads us to suggest the default value of $a_0 = 0.5$ in the Beta prior, in connection with both the earlier result of Rousseau and Mengersen (2011) and Jeffreys' prior in the simplest mixture setting.

A last point about our proposal is that it does not induce additional computational strain on the analysis. Provided algorithmic solutions exist for both models under comparison, such solutions can be recycled towards estimating the encompassing mixture model. As demonstrated through the various examples in the paper, the setting is actually easier than with a standard mixture estimation problem (Diebolt and Robert, 1994; Marin et al., 2005) because of the existence of common parameters that allow for the original MCMC samplers to be turned into proposals. Gibbs sampling completions are useful for assessing the potential outliers in a model but are not essential to achieve a conclusion about the overall problem.

REFERENCES

- ADAMS, M. (1987). *William Ockham*. University of Notre Dame Press, Notre Dame, Indiana.
 AITKIN, M. (1991). Posterior Bayes factors (with discussion). *J. Royal Statist. Society Series B*, **53** 111–142.

- AITKIN, M. (2010). *Statistical Inference: A Bayesian/Likelihood approach*. CRC Press, Chapman & Hall, New York.
- ATKINSON, A. (1970). A method for discriminating between models. *J. Royal Statist. Society Series B*, **32** 211–243.
- BALASUBRAMANIAN, V. (1997). Statistical inference, Occam’s razor, and statistical mechanics on the space of probability distributions. *Neural Computat.*, **9** 349–368.
- BAYARRI, M. and GARCIA-DONATO, G. (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, **94** 135–152.
- BERGER, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer-Verlag, New York.
- BERGER, J. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18** 1–32.
- BERGER, J., BOUKAI, B. and WANG, Y. (1997). Unified frequentist and Bayesian testing of a precise hypothesis (with discussion). *Statistical Science*, **12** 133–160.
- BERGER, J., BOUKAI, B. and WANG, Y. (1999). Simultaneous Bayesian-frequentist sequential testing of nested hypotheses. *Biometrika*, **86** 79–92.
- BERGER, J., GHOSH, J. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *Journal of Statistical Planning and Inference*, **112** 241–258.
- BERGER, J. and JEFFERYS, W. (1992). Ockham’s razor and Bayesian analysis. *Amer. Scientist*, **80** 64–72.
- BERGER, J. and PERICCHI, L. (1996). The intrinsic Bayes factor for model selection and prediction. *J. American Statist. Assoc.*, **91** 109–122.
- BERGER, J. and PERICCHI, L. (2001). Objective Bayesian methods for model selection: introduction and comparison. In *Model Selection* (P. Lahiri, ed.), vol. 38 of *Lecture Notes – Monograph Series*. Institute of Mathematical Statistics, Beachwood Ohio, 135–207.
- BERGER, J., PERICCHI, L. and VARSHAVSKY, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, **60** 307–321.
- BERGER, J. and SELLKE, T. (1987). Testing a point-null hypothesis: the irreconcilability of significance levels and evidence (with discussion). *J. American Statist. Assoc.*, **82** 112–122.
- BERKHOF, J., VAN MECHELEN, I. and GELMAN, A. (2003). A Bayesian approach to the selection and testing of mixture models. *Statistica Sinica*, **13** 423–442.
- BERNARDO, J. (1980). A Bayesian analysis of classical hypothesis testing. In *Bayesian Statistics* (J. Bernardo, M. H. DeGroot, D. V. Lindley and A. Smith, eds.). Oxford University Press.
- CARLIN, B. and CHIB, S. (1995). Bayesian model choice through Markov chain Monte Carlo. *J. Royal Statist. Society Series B*, **57** 473–484.
- CASELLA, G. and BERGER, R. (1987). Reconciling Bayesian and frequentist evidence in the one-sided testing problem. *J. American Statist. Assoc.*, **82** 106–111.
- CELEUX, G., HURN, M. and ROBERT, C. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. American Statist. Assoc.*, **95**(3) 957–979.
- CHEN, M., SHAO, Q. and IBRAHIM, J. (2000). *Monte Carlo Methods in Bayesian Computation*. Springer-Verlag, New York.
- CHIB, S. (1995). Marginal likelihood from the Gibbs output. *J. American Statist. Assoc.*, **90** 1313–1321.
- CHOPIN, N. and ROBERT, C. (2010). Properties of nested sampling. *Biometrika*, **97** 741–755.
- CHOUDHURY, A., RAY, S. and SARKAR, P. (2007). Approximating the cumulative distribution function of the normal distribution. *Journal of Statistical Research*, **41** 59–67.
- CHRISTENSEN, R., JOHNSON, W., BRANSCUM, A. and HANSON, T. (2011). *Bayesian ideas and data analysis: an introduction for scientists and statisticians*. CRC Press, New York.
- CONSONNI, G., FORSTER, J. J. and LA ROCCA, L. (2013). The whetstone and the alum block: Balanced objective Bayesian comparison of nested models for discrete data. *Statistical Science*, **28** 398–423.
- CSISZÁR, I. and SHIELDS, P. (2000). The consistency of the BIC Markov order estimator. *Ann. Statist.*, **28** 1601–1619.
- DE SANTIS, F. and SPEZZAFERRI, F. (1997). Alternative Bayes factors for model selection. *Canadian J. Statist.*, **25** 503–515.
- DEGROOT, M. (1970). *Optimal Statistical Decisions*. McGraw-Hill, New York.
- DEGROOT, M. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *J. American Statist. Assoc.*, **68** 966–969.
- DEGROOT, M. (1982). Discussion of Shafer’s ‘Lindley’s paradox’. *J. American Statist. Assoc.*, **378** 337–339.
- DICKEY, J. and GUNEL, E. (1978). Bayes factors from mixed probabilities. *Journal of the Royal Statistical Society. Series B (Methodological)*, **40** 43–46.
- DIEBOLT, J. and ROBERT, C. (1994). Estimation of finite mixture distributions by Bayesian sampling. *J. Royal Statist. Society Series B*, **56** 363–375.
- FERNANDEZ, C., LEY, E. and STEEL, M. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics*, **100** 381–427.

- FRÜHWIRTH-SCHNATTER, S. (2006). *Finite Mixture and Markov Switching Models*. Springer-Verlag, New York, New York.
- GELMAN, A. (2008). Objections to Bayesian statistics. *Bayesian Analysis*, **3**(3) 445–450.
- GELMAN, A., CARLIN, J., STERN, H., DUNSON, D., VEHTARI, A. and RUBIN, D. (2013a). *Bayesian Data Analysis*. 3rd ed. Chapman and Hall, New York, New York.
- GELMAN, A., ROBERT, C. and ROUSSEAU, J. (2013b). Inherent difficulties of non-Bayesian likelihood-based inference, as revealed by an examination of a recent book by Aitkin (with a reply from the author). *Statistics & Risk Modeling*, **30** 1001–1016.
- GEWEKE, J. (2010). *Complete and Incomplete Econometric Models*. The Econometric and Tinbergen Institutes lectures, Princeton University Press, Princeton, NJ.
- GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.*, **28** 500–531.
- GIGERENZER, G. (1991). The superego, the ego and the id in statistical reasoning. In *Methodological and Quantitative Issues in the Analysis of Psychological Data* (G. Keren and C. Lewis, eds.). Erlbaum, Hillsdale, New Jersey.
- GOURIÉROUX, C. and MONFORT, A. (1996). *Statistics and Econometric Models*. Cambridge University Press, Cambridge, UK.
- GREEN, P. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika*, **82** 711–732.
- HOETING, J. A., MADIGAN, D., RAFTERY, A. and VOLINSKY, C. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science*, **14** 382–417.
- JASRA, A., HOLMES, C. and STEPHENS, D. (2005). Markov Chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statist. Sci.*, **20** 50–67.
- JEFFERYS, W. and BERGER, J. (1992). Ockham’s razor and Bayesian analysis. *American Scientist*, **80** 64–72.
- JEFFREYS, H. (1939). *Theory of Probability*. 1st ed. The Clarendon Press, Oxford.
- JOHNSON, V. (2013a). Revised standards for statistical evidence. *Proc Natl Acad Sci USA*. URL <http://www.pnas.org/content/early/2013/10/28/1313476110.abstract>.
- JOHNSON, V. (2013b). Uniformly most powerful Bayesian tests. *J. Royal Statist. Society Series B*, **41** 1716–1741.
- JOHNSON, V. and ROSSELL, D. (2010). On the use of non-local prior densities in Bayesian hypothesis tests. *J. Royal Statist. Society Series B*, **72** 143–170.
- KASS, R. and RAFTERY, A. (1995). Bayes factors. *J. American Statist. Assoc.*, **90** 773–795.
- LAD, F. (2003). Appendix: the Jeffreys–Lindley paradox and its relevance to statistical testing. In *Conference on Science and Democracy, Palazzo Serra di Cassano, Napoli*.
- LAVINE, M. and SCHERVISH, M. J. (1999). Bayes factors: What they are and what they are not. *American Statist.*, **53** 119–122.
- LEE, K., MARIN, J.-M., Mengersen, K. and ROBERT, C. (2009). Bayesian inference on mixtures of distributions. In *Perspectives in Mathematical Sciences I: Probability and Statistics* (N. N. Sastry, M. Delampady and B. Rajeev, eds.). World Scientific, Singapore, 165–202.
- LINDLEY, D. (1957). A statistical paradox. *Biometrika*, **44** 187–192.
- MACKEY, D. J. C. (2002). *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, Cambridge, UK.
- MADIGAN, D. and RAFTERY, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *J. American Statist. Assoc.*, **89** 1535–1546.
- MARIN, J., PILLAI, N., ROBERT, C. and ROUSSEAU, J. (2014). Relevant statistics for Bayesian model choice. *J. Royal Statist. Soc. Series B*, **76** 833–859.
- MARIN, J. and ROBERT, C. (2007). *Bayesian Core*. Springer-Verlag, New York.
- MARIN, J. and ROBERT, C. (2011). Importance sampling methods for Bayesian discrimination between embedded models. In *Frontiers of Statistical Decision Making and Bayesian Analysis* (M.-H. Chen, D. Dey, P. Müller, D. Sun and K. Ye, eds.). Springer-Verlag, New York.
- MARIN, J.-M., Mengersen, K. and ROBERT, C. (2005). Bayesian modelling and inference on mixtures of distributions. In *Handbook of Statistics* (C. Rao and D. Dey, eds.), vol. 25. Springer-Verlag, New York, 459–507.
- MAYO, D. and COX, D. (2006). Frequentist statistics as a theory of inductive inference. In *Optimality: The Second Erich L. Lehmann Symposium* (J. Rojo, ed.). Lecture Notes-Monograph Series, Institute of Mathematical Statistics, Beachwood, Ohio, USA, 77–97.
- MCVINISH, R., ROUSSEAU, J. and Mengersen, K. (2009). Bayesian goodness-of-fit testing with mixtures of triangular distributions. *Scandinavian Journ. Statist.*, **36** 337–354.
- NEAL, R. (1994). Contribution to the discussion of “Approximate Bayesian inference with the weighted likelihood bootstrap” by Michael A. Newton and Adrian E. Raftery. *J. Royal Statist. Society Series B*, **56** (1) 41–42.
- NEAL, R. (1999). Erroneous results in “Marginal likelihood from the Gibbs output”. Tech. rep., University of

- Toronto. URL <http://www.cs.utoronto.ca/~radford>.
- NEWTON, M. and RAFTERY, A. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *J. Royal Statist. Society Series B*, **56** 1–48.
- NEYMAN, J. and PEARSON, E. (1933). The testing of statistical hypotheses in relation to probabilities a priori. *Proc. Cambridge Philos. Soc.*, **24** 492–510.
- O’HAGAN, A. (1995). Fractional Bayes factors for model comparisons. *J. Royal Statist. Society Series B*, **57** 99–138.
- O’NEILL, P. D. and KYPRAIOS, T. (2014). Bayesian model choice via mixture distributions with application to epidemics and population process models. *ArXiv e-prints*. [1411.7888](https://arxiv.org/abs/1411.7888).
- PESARAN, M. and DEATON, A. (1978). Testing non-nested nonlinear regression models. *Econometrica*, **46** 677–694.
- PLUMMER, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9** 523–539.
- QUANDT, R. (1974). A comparison of methods for testing nonnested hypotheses. *Review of Economics and Statistics*, **LVI** 92–99.
- R DEVELOPMENT CORE TEAM (2006). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <http://www.R-project.org>.
- RASMUSSEN, C. E. and GHAHRAMANI, Z. (2001). Occam’s razor. In *Advances in Neural Information Processing Systems*, vol. 13.
- RICHARDSON, S. and GREEN, P. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B*, **59** 731–792.
- ROBERT, C. (1993). A note on Jeffreys–Lindley paradox. *Statistica Sinica*, **3** 601–608.
- ROBERT, C. (2001). *The Bayesian Choice*. 2nd ed. Springer-Verlag, New York.
- ROBERT, C. (2014). On the Jeffreys–Lindley paradox. *Philosophy of Science*, **5** 216–232.
- ROBERT, C., CHOPIN, N. and ROUSSEAU, J. (2009). Theory of Probability revisited (with discussion). *Statist. Science*, **24**(2) 141–172 and 191–194.
- ROBERT, C., CORNUET, J.-M., MARIN, J.-M. and PILLAI, N. (2011). Lack of confidence in ABC model choice. *Proceedings of the National Academy of Sciences*, **108**(37) 15112–15117.
- RODRIGUEZ, C. and WALKER, S. (2014). Label switching in Bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, **23** 25–45.
- ROUSSEAU, J. (2007). Approximating interval hypotheses: p-values and Bayes factors. In *Bayesian Statistics 8: Proceedings of the Eighth International Meeting* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.). Oxford University Press.
- ROUSSEAU, J. and MENGENSEN, K. (2011). Asymptotic behaviour of the posterior distribution in overfitted mixture models. *J. Royal Statist. Society Series B*, **73** 689–710.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6** 461–464.
- SHAFFER, G. (1982). On Lindley’s paradox (with discussion). *Journal of the American Statistical Association*, **378** 325–351.
- SKILLING, J. (2006). Nested sampling for general Bayesian computation. *Bayesian Analysis*, **1**(4) 833–860.
- SPANOS, A. (2013). Who should be afraid of the Jeffreys–Lindley paradox? *Philosophy of Science*, **80** 73–93.
- SPIEGELHALTER, D. J., BEST, N., B.P., C. and VAN DER LINDE, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, **64** 583–640.
- SPRENGER, J. (2013). Testing a precise null hypothesis: The case of Lindley’s paradox. *Philosophy of Science*, **80** 733–744.
- STEELE, R., RAFTERY, A. and EMOND, M. (2006). Computing normalizing constants for finite mixture models via incremental mixture importance sampling (imis). *Journal of Computational and Graphical Statistics*, **15** 712–734.
- STEPHENS, M. (2000). Dealing with label switching in mixture models. *J. Royal Statist. Society Series B*, **62**(4) 795–809.
- VAN HAVRE, Z., MENGENSEN, K., ROUSSEAU, J. and WHITE, N. (2014). Addressing open questions in mixture models. Tech. Rep. 1412.08, QUT, Department of Statistics, Technical Report Series.
- VEHTARI, A. and LAMPINEN, J. (2002). Bayesian model assessment and comparison using crossvalidation predictive densities. *Neural Computation*, **14** 2439–2468.
- VEHTARI, A. and OJANEN, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6** 142–228.
- ZELLNER, A. (1986). On assessing prior distributions and Bayesian regression analysis with *g*-prior distribution regression using Bayesian variable selection. In *Bayesian inference and decision techniques: Essays in Honor of Bruno de Finetti*. North-Holland / Elsevier, 233–243.
- ZILIAK, S. and MCCLOSKEY, D. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Univ of Michigan Pr.

APPENDIX 1: MARGINAL LIKELIHOOD FOR THE LAPLACE DISTRIBUTION

In order to derive the Bayes factor in Example 3.4, we compute here the marginal likelihood of the double-exponential $\mathcal{L}(\mu, \sqrt{2})$ model under a flat prior, as follows:

$$\begin{aligned} \int_{-\infty}^{\infty} \exp \left\{ -1/\sqrt{2} \sum_{i=1}^n |x_i - \mu| \right\} d\mu &= \sqrt{2}/n \exp \left\{ -1/\sqrt{2} \left(\sum_{j=1}^n x_{(j)} - nx_{(1)} \right) \right\} \\ &+ \sqrt{2} \sum_{\substack{i=1 \\ i \neq n/2}}^{n-1} 1/n-2i \exp \left\{ -1/\sqrt{2} \left(\sum_{j=i+1}^n x_{(j)} - \sum_{j=1}^i x_{(j)} - (n-2i)x_{(i+1)} \right) \right\} \\ &- \sqrt{2} \sum_{\substack{i=1 \\ i \neq n/2}}^{n-1} 1/n-2i \exp \left\{ -1/\sqrt{2} \left(\sum_{j=i+1}^n x_{(j)} - \sum_{j=1}^{i-1} x_{(j)} - (n-2i+1)x_{(i)} \right) \right\} \\ &+ (x_{n/2+1} - x_{n/2}) \exp \left\{ -1/\sqrt{2} \left(\sum_{j=\frac{n}{2}+1}^n x_{(j)} - \sum_{j=1}^{\frac{n}{2}} x_{(j)} \right) \right\} \\ &+ \sqrt{2}/n \exp \left\{ -1/\sqrt{2} \left(nx_{(n)} - \sum_{j=1}^n x_{(j)} \right) \right\} \end{aligned}$$

where $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ denotes the order statistics.

APPENDIX 2: PROPRIETY OF THE POSTERIOR IN THE CASE STUDY OF SECTION 4

To prove the propriety of the posterior it is enough to prove the propriety of the sub-posterior distribution associated to each component since the parameter (θ, σ) is shared between the components. It is known that in the case of a Gaussian model $\mathcal{N}(\theta, \sigma)$ the posterior associated to the prior $\pi(\theta, \sigma) = 1/\sigma$ is proper as soon as $n \geq 2$ and at least 2 observations are distinct. We now show that this result extends to the case of a Gumbel (θ, σ^2) and of a Logistic (θ, σ) . Let I denote the marginal likelihood, in the Gumbel case.

$$I = \int_{\mathbb{R} \times \mathbb{R}^+} \frac{1}{\sigma^{n+1}} \exp \left\{ -\sum_{i=1}^n (Y_i - \theta)/\sigma \right\} \exp \left\{ -\sum_{i=1}^n e^{-(Y_i - \theta)/\sigma} \right\} d\theta d\sigma$$

We set $\gamma_n = \sum_{i=1}^n e^{-Y_i/\sigma}$, then

$$\begin{aligned} I(\sigma) &= e^{-n\bar{Y}_n/\sigma} \int_{\mathbb{R}} \exp(n\theta/\sigma) \exp(-\gamma_n e^{\theta/\sigma}) d\theta \propto e^{-n\bar{Y}_n/\sigma} \sigma \int_{\mathbb{R}} u^{n-1} \exp(-\gamma_n u) du \\ &\propto e^{-n\bar{Y}_n/\sigma} \sigma \gamma_n^{-n} \propto \exp \left(-\frac{n\bar{Y}_n}{\sigma} \right) \left(\sum_{i=1}^n e^{-\frac{Y_i}{\sigma}} \right)^{-n}. \end{aligned}$$

So

$$I \propto \int_{\mathbb{R}^+} \sigma^{-n} \exp \left(-\frac{n\bar{Y}_n}{\sigma} \right) \left(\sum_{i=1}^n e^{-\frac{Y_i}{\sigma}} \right)^{-n} d\sigma = \int_{\mathbb{R}^+} \sigma^{-n} \left(\sum_{i=1}^n e^{-\frac{1}{\sigma}(Y_i - \bar{Y}_n)} \right)^{-n} d\sigma < +\infty$$

if only if $\min_i (Y_i - \bar{Y}_n) < 0$. This is almost surely true when $n \geq 2$. We now study the Logistic case, using similar computations, so that

$$\begin{aligned}
I &\propto \int_{\mathbb{R} \times \mathbb{R}^+} \frac{e^{-n \frac{\bar{Y}_n}{\sigma}}}{\sigma^{(n+1)}} \frac{e^{\theta n / \sigma}}{\prod_i (1 + e^{-Y_i / \sigma} e^{\theta / \sigma})^2} d\theta d\sigma \propto \int_{\mathbb{R}^+} \frac{e^{-n \frac{\bar{Y}_n}{\sigma}}}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{\prod_i (1 + e^{-Y_i / \sigma} u)^2} du d\sigma \\
&\leq \int_{\mathbb{R}^+} \frac{e^{-n \frac{\bar{Y}_n}{\sigma}}}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{(1 + u^{n-1} e^{-(n-1)\bar{Y}_n / \sigma} \max_i e^{-(Y_i - \bar{Y}_n) / \sigma})^2} du \\
&\propto \int_{\mathbb{R}^+} \frac{1}{\sigma^n} \int_{\mathbb{R}} \frac{u^{n-1}}{(1 + u^{n-1} \max_i e^{-(Y_i - \bar{Y}_n) / \sigma})^2} du \\
&\propto \int_{\mathbb{R}^+} \frac{1}{\sigma^n} e^{2n \min_i (Y_i - \bar{Y}_n) / \sigma} d\sigma < +\infty
\end{aligned}$$

if and only if $\min_i (Y_i - \bar{Y}_n) < 0$. Thus means the observations cannot be all equal.